From Language to Location Using Multiple Instance Neural Networks

Sneha Nagpaul(✉) and Huzefa Rangwala

George Mason University, Fairfax, VA, USA
snagpaul@gmu.edu, rangwala@cs.gmu.edu

Abstract. Language patterns pertaining to a geographic region has various uses including cultural exploration, disaster response and targeted advertising. In this paper, we propose a method for geographically locating short text data within a multiple instance learning framework augmented by neural networks. Our representation learning approach tackles minimally pre-processed social media discourse and discovers high level language features that are used for classification. The proposed method scales and adapts to datasets relating to 15 cities in the United States. Empirical evaluation demonstrates that our approach outperforms state of the art in multiple instance learning while providing a framework that alleviates the need for subjective feature engineering based approaches.

Keywords: NLP · MIL · Text geolocation · Neural networks

1 Introduction

Due to privacy concerns, users of social media often chose not to share the geographic location while they generate content. Besides commercial and malicious uses such as targeted advertising and recommender systems [3], this information could also be used to facilitate better disaster response and help law enforcement [1] with crime prevention [8]. Thus, a system which geo-tags user generated text is valuable for its social applications.

Prior work on text geolocation frames the problem as classification of user discourse into regions based on words that appear in the text [9]. Since the phrase level structure is distorted by this Bag Of Words approach, these models often lose context because word order is lost. Additionally, the data requirements tend to move away from short text to body of text produced by a user. Hence, they end up predicting a user's location rather than locating a stand alone piece of content.

Hence, this is a problem where location is available for users rather than an individual tweet. This allows us to express the problem for distilling information from group level labels to individual parts within the group. This is referred to as multiple instance learning (MIL) and has found extensive use in semi-supervised learning and sentiment analysis. Since MIL research makes strong assumptions

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 46–53, 2018.
https://doi.org/10.1007/978-3-319-93372-6_5

about the membership of instances inside a bag and/or use feature engineering based approaches, there is scope to augment this work. This paper makes a contribution to the tractability and abstraction mechanisms employed within an MIL exercise.

The approach proposed in this work provides a flexible and scalable framework for transferring geographic location labels from user to tweet level without the use of explicit features or kernels. The flexibility is present in its modular structure that separates instance level predictions from bag level aggregations while still enabling a backward flow of information of labels from the aggregate to the individual instance level. Since theI underlying method is a neural network that can be trained using optimization techniques and learns internal representations in the datasets, it scales well to the size and types of datasets under consideration.

2 Related Work

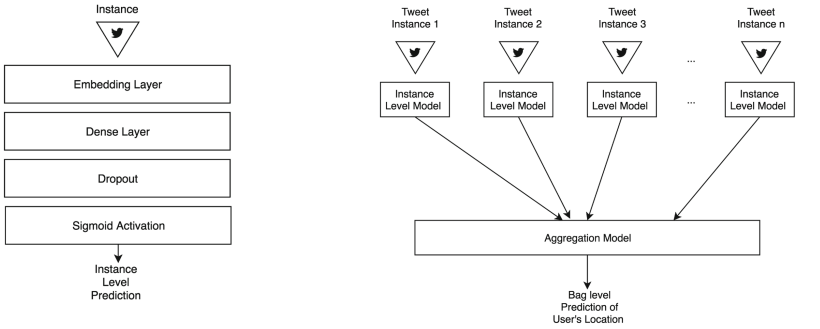
Multiple Instance Learning. Within the standard formulation, a group of instances referred as *bags* are labeled but individual instances are not [11] The bag level label is associated with its contents by a membership assumption and an aggregation function.

Single Instance Learning (SIL) is a naive and noisy way to accomplish this task [10] wherein every instance is assigned the label of its bag. Recently, neural networks have been used with adapted cost functions to accomplish the task of relaxing aggregation assumptions while using custom similarity measures [5]. Most of these prior methods are kernel based, as they require substantial feature engineering and are thus hard to scale. Additionally, in prior applications instances share heavy context, whereas tweets within a user-bag need not share context or even temporal origins.

Geographic Information Retrieval. Geographic Information Retrieval refers to methods that deal with mapping language to location [7]. Classically, a supplementary dataset or gazetteer, that maps words to locations along with heuristics to disambiguate place names was used. However as scale of datasets grew language modeling became prevalent in GIR which solves the problem for the user level with Bag of Words models using traditional bayesian techniques and using neural networks [7].

3 Methods

To overcome the gaps in prior work, we leverage basic feed forward neural network architectures like the multi layer perceptron (MLP) [6]. We make simple changes to this basic architecture to enable it to perform MIL with higher level of modular abstraction for instance level classification and bag level aggregation, as shown in Fig. 1b. We consider a user-bag, labeled with a binary location label which contains tweet-instances that are devoid of labels at the training stage.



(a) Instance Level Model

(b) Model Architecture

Fig. 1. (a): Instance level model: the model for each tweet and (b) model architecture: to achieve MIL, the instance level models feed their predictions to a bag level aggregation layer to be able to share the weights from retro-propagated losses.

3.1 Problem Statement

Given a user U_i with a binary location label $y_i \in \{0, 1\}$ where 1 denotes that the user is from a particular city and 0 denotes otherwise. Each U_i is a collection of tweets $t_{ij}, j = 1, 2, \ldots N$ and the task is then to devise a function $f(t) \rightarrow y$ which essentially labels individual tweets as belonging to the city under consideration.

For a treatment of the problem as formulated here, an end-to-end trainable neural network architecture is proposed in this work and is called milNN. The model's architecture is illustrated in Fig. 1a and b.

Instance Level Classifier. The tweet level classifier consists of an embedding layer that feeds into the fully connected hidden layer component and is designed as though labels were available (Fig. 1a). The embedding layer learns representations that can be viewed as an intuitive language model as opposed to a symbolic language model that stems from rigid grammatical rules or engineered features [6]. This also makes the model well suited to social media content which often deviates from traditional language use.

Bag Level Classifier. The instance level classifier is then applied to individual tweets and the results are averaged to get the bag level labels as shown in Fig. 1b. This is the component of the architecture that addresses the relationship between bag and instance level labels.

Loss Function and Training. At this stage a label is available and losses (binary cross entropy) can be back-propagated (Adam Optimizer [4]) throughout the network. Thus, the instance level classifier gets trained as a result of gradients of the bag level losses.

3.2 Method Characteristics

Due to being fully embedded in the neural network and representation learning paradigm, milNN relies on learning distributed representations and is devoid of subjective feature engineering requirements. This also equips it to handle any changes that might occur organically in the data. Moreover, this framework does not have high computational and memory requirements and learned using stochastic gradient descent which is easy to parallelize.

Additionally, the assumptions of membership proportions are relaxed and aggregation assumptions are not particularly stringent as the sigmoid layer does not provide exact labels for the instances, but rather a probabilistic average across all tweet classifications outputs for a user level label. The architecture is also flexible and the model described here can be seen as one example of the most basic possibilities.

The datasets were created by reverse geocoding information from Twitter North America dataset [12]. The latitude and longitude readings were recorded when the user registered on Twitter and provided the location. Subsequent tweets were recorded for this user. For use in this work, the top cities in the data were split into fifteen datasets of equal number of positive and negative samples. The negative samples for each city were randomly selected from the rest of the dataset after stratified sampling from other cities.

3.3 Experimental Setup

At the instance level, the tweets are preprocessed by changing URLs, @mentions, and hashtags to a generic tags for each. Subsequently they are tokenized and vocabulary size is chosen to be 5000. The tweet is then padded to a 20 word maximum and then fed through an embedding layer with 32 dimensions which is randomly initialized. Following this, there is a single hidden layer with 100 nodes that process the various language level relationships and feed the relu activations to the sigmoid layer for classification after adding a dropout of 25% for regularization. 10 tweets from each user are averaged from the instance model at a higher layer for the bag level output. At this level binary cross entropy loss is calculated using the bag level labels and back-propagated using the Adam optimizer. A batch size of 256 bags at a time is chosen and trained for 200 epochs with a learning rate of 0.0001. An early stopping condition is included which breaks out of training when the loss of the epoch converges and waits for 5 iterations to confirm the convergence. The hyper-parameters chosen for the model are described in this section and were chosen using half the training data for validation. The performance of all considered choices was comparable except for a running time increase for models with more parameters.

3.4 Results

As seen in Table 1, in terms of the accuracy metric milNN outperforms state of the art on 14 of the 15 datasets considered here. When considering the F-score, it outperforms the other methods on 10 of the 15 datasets. It is important to

notice that when it loses to older methods, it is for smaller datasets and not by a lot of margin. However, when it outperforms it is significantly better (eg. Boston performance is better by 20%). Also, it is consistently good on datasets of varied sizes and needs while the other methods don't seem to be able to adapt to the feature requirements and scale of the data.

Table 1. Accuracy and F-scores for milNN and prior methods. milNN scores a 14/15 and 10/15 on Accuracy and F-score respectively on the 15 datasets of varied sizes (train-test split was 80:20)

City	Acc:SIL	Acc:GICF	Acc:milNN	F1:SIL	F1:GICF	F1:milNN	Total
Atlanta	0.5780	0.6025	**0.6602**	0.6900	**0.6982**	0.6568	4414
Austin	0.6070	0.6501	**0.7015**	0.6650	0.6291	**0.6848**	2915
Baltimore	0.5180	0.6248	**0.6858**	0.6560	**0.6957**	0.6816	2700
Boston	0.5460	0.5774	**0.6276**	0.5820	0.5960	**0.6130**	2389
Chicago	0.5760	0.6429	**0.6502**	0.5180	0.5163	**0.6420**	8286
New Orleans	0.5230	0.6365	**0.6962**	0.6690	0.6976	**0.7041**	2592
New York City	0.5740	0.6476	**0.7024**	0.6230	0.6349	**0.6988**	19000
Paradise	0.6060	**0.6629**	0.6565	**0.6510**	0.6365	0.6468	3095
Philadelphia	0.5190	0.6195	**0.6644**	0.6620	**0.7006**	0.6830	5792
San Diego	0.6300	0.6504	**0.6850**	0.6080	0.6325	**0.6548**	2452
San Francisco	0.6300	0.7322	**0.7542**	0.6980	0.7398	**0.7603**	7710
Seattle	0.6210	0.6970	**0.7269**	0.6840	0.7112	**0.7215**	3350
Toronto	0.6390	0.6895	**0.7520**	0.6810	0.7056	**0.7485**	5037
Washington, D.C.	0.5740	0.6298	**0.6437**	**0.6190**	0.4910	0.6108	5732
Weehawken	0.6160	0.6727	**0.7000**	0.6590	0.6588	**0.6827**	2196

3.5 Case Study

Since San Francisco was the biggest dataset and milNN outperformed the state of the art on all counts due to its scalability. The structure of the following analysis is to go over examples of users that belong to SF for exploring language patterns. Additionally the anti-patterns are explored by analyzing tweets from users that do not belong in order to discover how SF users don't tweet.

The word cloud visual is created using all the test instance tweets that were classified to be over 0.95 by the instance level classifier. Location entities were subsequently extracted from these tweets using StandfordNER [2] and then weighted into a word cloud (Fig. 2).

While the San Francisco model caught traditional place names such as 'Alcatraz' and 'San Franciso', more exotic language patterns were also discovered in this dataset. For the second user that was from SF, technology related tweets were identified as indicative of the region. Additionally, a proclivity to tweet with correct grammar is also discovered when a single word change causes probability to increase as grammatical usage becomes less awkward. This point is further reinforced when a user that is not from SF is seen to have no high probability tweets due to use of slang (Fig. 3).

Fig. 2. Word cloud - SF

San Francisco - User 1

0.9999975- " I'm at Alcatraz (Alcatraz Island, San Francisco Bay, San Francisco) w/ 6 others http://t.co/47YWmX9p "

0.9999856 - " I'm at Chinatown Gate (500 Bush St, at Grant Ave, San Francisco) http://t.co/rb49RnFa "

0.5055542 - " @matthewharkin @phillo haha, now I'm worried ")

0.025480814 - " I'm at Tiffany & Co. (210 N Rodeo Dr., Beverly Hills) http://t.co/YppqS7ix ")

San Francisco - User 2

0.99910492 - "Can't wait for @BankSimple, @usbank is such a joke from a technology / ease-of-use perspective."

0.54251802 - "My whole morning **has** been devoted to banking. Not done yet. Living the life."

0.3358801 - "My whole morning **had** been devoted to banking. Not done yet. Living the life."

User not from San Francisco

0.16304019 - "Follow the OG triple OG @thad4mayor to ensure that he don't steal ur wallet when he see you in the streets...<>jtfo"

0.23261635 - "@jbdachamp u show me no luv :("

0.0011654327 - "Nap time"

0.011714808 - "somethins gotta give"

0.10918618- "@Cree_Oh_Lay_CO how ya been?"

0.00020607341 -"@jbdachamp and u won't lol",

0.0094076423 - "@jbdachamp I was MIA 4 a min due 2 **technical** issues but now I'm baaaaack lol"

0.010172283 - "da best part is that the downs dont last always"

0.20761815 - "I luv fridays :)"

0.064976566 - "TGIF"

0.073392898 - "@Cree_Oh_Lay_CO we're great :)"

0.18137941 - "@thad4mayor "our" hmmm lol"

Fig. 3. San Francisco examples

4 Conclusion

This work demonstrated a flexible multiple instance learning framework applied to identifying geographic location of short text data. The experiments showed that milNN was scalable and capable of discovering high level language features such as grammar in addition to place names in data. Thus this work contributed to multiple instance learning and geographic information retrieval literature by designing a novel model architecture that was end-to-end trainable.

While recurrent neural networks and attention mechanisms might have lent themselves to the problem considered here, we chose to focus this exercise on MIL in order to augment prior research that has suffered from the intractability of kernel based methods. Given the flexibility of the neural network architecture, future work could focus on developing recurrent neural network based models that can take a variable length input both in terms of tweet length and number of tweets. Additionally, extensions of standard aggregation functions could be developed for instance level data that have bag-internal structure that needs to be exploited, like in reviews. Transfer learning approaches could also be explored given the embedding based input of the neural network as and when reliable twitter word vectors become available.

Acknowledgment. This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2017-ST-061-CINA01. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

References

1. Ashktorab, Z., Brown, C.D., Nandi, M., Culotta, A.: Tweedr: mining Twitter to inform disaster response. In: ISCRAM (2014)
2. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL 2005, pp. 363–370. Association for Computational Linguistics, Stroudsburg (2005). https://doi.org/10.3115/1219840.1219885
3. Ho, S.S., Lieberman, M., Wang, P., Samet, H.: Mining future spatiotemporal events and their sentiment from online news articles for location-aware recommendation system. In: Proceedings of the First ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems, MobiGIS 2012, pp. 25–32. ACM, New York (2012). http://doi.acm.org/10.1145/2442810.2442816
4. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. CoRR abs/1412.6980 (2014). http://arxiv.org/abs/1412.6980
5. Kotzias, D., Denil, M., de Freitas, N., Smyth, P.: From group to individual labels using deep features. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2015, pp. 597–606. ACM, New York (2015). http://doi.acm.org/10.1145/2783258.2783380
6. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–44 (2015)
7. Melo, F., Martins, B.: Automated geocoding of textual documents: a survey of current approaches. Trans. GIS **21**(1), 3–38 (2016). https://doi.org/10.1111/tgis.12212
8. Ning, Y., Muthiah, S., Rangwala, H., Ramakrishnan, N.: Modeling precursors for event forecasting via nested multi-instance learning. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016, pp. 1095–1104. ACM, New York (2016). http://doi.acm.org/10.1145/2939672.2939802

9. Rahimi, A., Cohn, T., Baldwin, T.: A neural model for user geolocation and lexical dialectology. CoRR abs/1704.04008 (2017). http://arxiv.org/abs/1704.04008
10. Ray, S., Craven, M.: Supervised versus multiple instance learning: an empirical comparison. In: Proceedings of the 22nd International Conference on Machine Learning, ICML 2005, pp. 697–704. ACM, New York (2005). http://doi.acm.org/10.1145/1102351.1102439
11. Foulds, J., Frank, E.: A review of multi-instance learning assumptions. Knowl. Eng. Rev. **25**(1), 1–25 (2010)
12. Roller, S., Speriosu, M., Rallapalli, S., Wing, B., Baldridge, J.: Supervised text-based geolocation using language models on an adaptive grid. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, pp. 1500–1510. Association for Computational Linguistics, Stroudsburg (2012). http://dl.acm.org/citation.cfm?id=2390948.2391120