GeoAttn: Localization of Social Media Messages via Attentional Memory Network

Sha Li[1], Chao Zhang[2], Dongming Lei[1], Ji Li[1], and Jiawei Han[1]

[1]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL USA
[2]College of Computing, Georgia Institute of Technology, Atlanta, GA USA
[1] {shal2, czhang82, dlei5, jili3, hanj}@illinois.edu

Abstract

Recent studies have demonstrated inspiring success in leveraging geo-tagged social media data for applications such as event detection, location recommendation and mobile healthcare. However, in most real-life social media streams, only a small percentage of data have explicit geo-location metadata, which hinders the power of social media from being fully unleashed.

We study the problem of inferring geo-locations from social media messages. While a number of text-based geo-locating techniques have been proposed, they either fall short of automatically identifying indicative keywords from noisy social media posts or do not integrate rich prior knowledge of geological regions. We propose an attentive memory network called GeoAttn for localization of social media messages. To capture indicative keywords for location inference, GeoAttn consists of an attentive message encoder, which selectively focuses on location-indicative terms to derive a discriminative message representation. The message embedding is then fed into a memory network, which selectively attends to relevant Points-of-Interest (POIs) for location prediction. The message encoder and key-value memory network are jointly trained in an end-to-end manner. The attention mechanisms in GeoAttn not only alleviate noisy information for higher prediction accuracy, but also provide interpretable attention scores that rationalize the predictions. Our experiments on a million-scale geo-tagged tweet dataset show that GeoAttn outperforms previous state-of-the-art location prediction methods by 15.5% in mean error distance, and is capable of locating over half of the tweets within 5km.

1 Introduction

The location information contained in social media data enables linking people's online posts to their physical-world activities, and plays an important role in intelligent location-based systems. For example, recent studies have demonstrated inspiring success in leveraging geo-tagged social media for a wide range of applications including data-driven traffic scheduling[10], urban planning[15], event detection[39], POI recommendation[40, 41] and personal healthcare[20, 30]. However, in a typical social media stream (*e.g.*, Twitter), only less than 1% records are associated with explicit GPS information. The *localization problem*—which aims at inferring the locations of social media messages—has thus become an important issue for unlocking the potential of social media and building intelligent location-based systems.

Earlier attempts to this problem are mostly *gazetteer-based* [12, 19], maintaining a look-up table from location entity names to real-world geographical locations. Such gazetteer-based methods are heavily limited by the scope and accuracy of the used gazettes. They also have difficulty in handling aliases and abbreviations, both of which are abundant in social media streams. Extensions of *topic models* [1, 4, 7, 28, 38] to jointly model geo-location and text have also been used for location prediction. The performance of such models is largely limited by the assumptions they make regarding the distribution of location-indicate keywords. Recently, a series of *classification methods* [27, 34, 35] have been purposed and have shown to produce the state-of-the-art performance for text localization. These models directly cast the localization problem as a classification task on geodesic grids but how to select such grids pose a challenge on their own. There are other works that take advantage of information beyond textual messages, such as social network relations and message metadata to predict the location of the user[6, 11, 29]. These methods are largely orthogonal to ours.

Linking messages to the correct locations faces two major challenges. The first is to identify location indicative keywords from notoriously short and noisy social media text. Current state-of-the-art methods mainly rely on preprocessing to remove stopwords and

normalize abbreviations. The remaining keywords are then treated equally. However, this is counter-intuitive for social media posts: for a post "DTW is closed because of freezing rain! Flight delayed twice then cancelled", the words "DTW" (referring to Detroit Metro Airport) and "flight" are obviously more useful in the location prediction task than "rain" or "delay". The second challenge is to leverage the existing rich prior knowledge of regions. In the localization task, it is often mentions of places, events and activities that differentiate one area from another. This is an excellent opportunity to leverage prior knowledge for useful metadata and exploit semantic connections between activities and their venues. By simply casting the localization problem as classification over geodesic grids, the semantic aspect of regions are overlooked.

Contributions. We propose GeoAttn for prediction of geo-locations of social media messages. At the high level, GeoAttn jointly learns the location aspect representation for messages and POI-anchored regions to encode their semantics, and performs localization by comparing the encoded message to region representation. In essence, we treat the problem as cross-modal matching instead of classification over grid-like areas. The whole model is end-to-end trainable without the need to manually assign weights to keywords. Moreover, the attention scores over keywords and POIs offer intuitive explanations that rationalize the prediction process.

To realize this goal, GeoAttn features two important modules: (1) an attentional message encoder; and (2) a key-value memory network[31]. Built upon a recurrent neural network, the message encoder derives a discriminative message representation by modeling the word sequence and selectively attending to the keywords that are location-indicative. To map keywords to geographical locations, we employ a key-value memory network. During prediction, we use the message representation from the message encoder to apply a soft attention layer over all entries in order to output a probability distribution over geographic space.

We highlight the contributions of this paper as follows:

1. We propose an attentional memory network framework for localization of social media messages. The framework bridges the text and location modalities by a key-value memory structure, and is capable of leveraging existing POI knowledge to facilitate accurate location prediction.
2. We design attention mechanisms over both the messages and regions. The attention mechanisms not only alleviate the effect of noisy information, but also offer interpretable explanations of the prediction process.
3. We have performed extensive experiments on million-scale tweet datasets. Our experimental results show that GeoAttn reduces the mean error distance by more than 15.5% compared to the best-performing baseline. Furthermore, the derived attention scores are highly meaningful in terms of assigning messages into proper locations.

2 Related Work

2.1 Geolocation Prediction Existing studies have investigated geolocation prediction at two different levels: *user localization* and *document localization*.

User localization aims at predicting the home location of social network users. The prediction granularity varies from city level to state level or even country level [27]. Based on the data used, there are three lines of approaches for user localization: text-based [4, 9, 14, 26, 28, 34], network-based [6, 11, 29] and a hybrid of the two [18, 23]. Existing text-based user localization methods predominantly cast the problem as a multi-class classification problem [9, 14, 26, 34]. Network-based approaches assume that friends in a social network are geographically close[6, 11, 29]. Hybrid approaches [18, 23, 25] combine knowledge from both text and networks for location prediction. The application of user-level localization is limited as users are treated as static throughout time which is necessary for mobility modeling, personalized recommendation, etc.

Document localization, which attempts to infer the geolocation of a specific document, is more closely related to our study. Geographic topic models [1, 7, 38] extend classic topic models by assuming each latent topic has distributions over not only textual keywords but also geographical coordinates. Supervised classification methods have also been applied to this problem using textual features[34]. Compared with these document localization methods, our model employs distributional representation of words to address the sparsity problem and utilizes the attention mechanism to perform automatic feature selection. Furthermore, we incorporate prior knowledge on POIs through the memory component, giving us better accuracy with less training data and also better intrepretability.

2.2 Attention Mechanisms and Memory Networks Attention mechanisms empower models with the ability to extract local features and assign different importance to different sections of the input[2]. Vaswani *et al.* [32] present a concise definition of attention as "mapping a query and a set of key-value pairs to an output". The attention mechanism has been widely adopted for deriving textual representation for tasks including machine translation [2], image captioning [37]

and visual question answering [36]. We are among the first to use the attention mechanism for the localization problem. In our model, the attention mechanism automatically selects words that are location-indicative and matches messages with location representations.

Memory networks [33] get their name from a long-term memory component that can be read and written to. Sukhbaatar *et al.* [31] proposed a continuous variant of the memory network that could be trained in end-to-end fashion. Miller *et al.* [17] demonstrate the flexibility of Key-Value Memory Networks in exploiting different knowledge sources. The key-value structure provides more possibilities in encoding prior knowledge and allows nontrivial transforms between keys and values. In our setting, the key-value memory network is used to bridge text and location.

3 Problem Description

We address the localization problem for individual messages in a supervised setting. Our input consists of two parts: a collection $\mathcal{C}$ of social media records $\{r_1, r_2, \ldots, r_M\}$ and auxiliary prior knowledge which is a collection $\mathcal{P}$ of regions or POIs $\{p_{n_1}, p_{n_2}, \ldots, p_{n_P}\}$.

Every record $r_m \in \mathcal{C}$ is a two-element tuple (m, l), where $l = (\phi, \lambda)$ is the location of record creation, and m is a piece of user-generated text.

Each of these known POIs p is also a two-element tuple (d, l) where d is its textual metadata including name, category and optionally description. l is the coordinates of the centroid of p.

Our goal is learn a model M on historical social media messages $\mathcal{C}$ with the help of the auxiliary data $\mathcal{P}$. When given a future test social media record (m, l), the model is expected to recover the ground-truth location l in terms of a distribution over geological space based on the text message m.

4 The GeoAttn Model

In this section, we first describe the overall design philosophy of GeoAttn, and then the details of different modules in GeoAttn.

4.1 Model Overview GeoAttn is designed for the message-level localization problem using only text data. Rather than discretizing geographic space using heuristics, we directly output a probability distribution estimate over *continuous space*.

Our model features automatic feature selection via the message attention layer, cross-modal translation with the help of POI metadata via the memory network and interpretable prediction results via the POI attention layer.

Concretely, as shown in Figure 1, GeoAttn consists of two major components: (1) the message encoder and (2) the POI memory network. With an attentional recurrent neural network (RNN), the message encoder generates a low-dimensional representation of the location-related aspect of message semantics. This message representation is then mapped to locations by the memory network through another attention layer. The resulting output is a probability distribution over the geographical area.

4.2 The Message Encoder The message encoder is designed to generate a low dimensional vector representation for each input message m, which is a variable length word sequence $\{u_1, u_2, \cdots u_n\}$. We design the message encoder as an attentional bidirectional RNN, detailed as follows.

4.2.1 Word Embeddings. Before feeding our message into the recurrent neural network, we map the words to low-dimensional embedding vectors.

Word embeddings allow us to generalize beyond symbolic matching and utilize semantic similarity. As shown in Figure 1, an embedding layer Φ is applied to map the input keywords $\{u_1, u_2, \cdots u_n\}$ in the message into a vector sequence. We use the *GloVe* model [22] to train word embeddings on the training set of our Twitter text corpus and make the embeddings fine-tunable. The training objective of GloVe is to learn word vectors such that their dot product equals the logarithm of the words' probability of co-occurrence.

We also use an existing POS tagger [21] to obtain POS tags for each word and append the one-hot encoding to the GloVe embeddings.

4.2.2 The Recurrent Unit. Word embeddings are used as input to a bi-directional RNN to derive a representation of the entire text message. The RNN preserves word order information and produces context-aware hidden representations for each word. We choose the gated recurrent unit (GRU) [5] due to its higher efficiency. From a length-n word embedding sequence, the RNN produces n hidden states $\{h_1, \cdots h_n\}$:

$$\{h_1, \cdots h_n\} = \text{GRU}(\{\Phi(u_1), \cdots \Phi(u_n)\}) \tag{4.1}$$

To further enhance the message representation, we make the GRU-based RNN bi-directional. Namely, in addition to feeding the original word sequence into the RNN, we also reverse the word sequence and feed it into another RNN. At each time step, we concatenate the latent states from both directions to form the representation at time step t, $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$.

4.2.3 The Attention Mechanism. Not all words in the message are equal: we wish to focus only on the words that are location-indicative, preferring POI name mentions, venue types and activity descriptions. To address this issue, we introduce an attention layer over the sequence of hidden states in the RNN. The attention layer performs automatic feature selection and

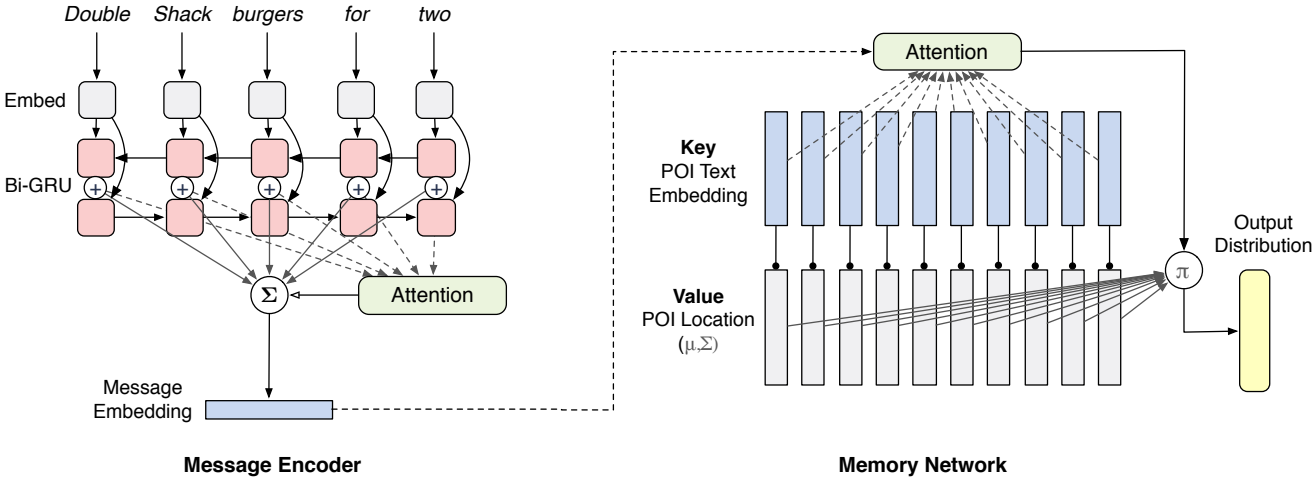

Figure 1: GeoAttn has two major components: (1) the message encoder and (2) the POI memory network. The message encoder generates low-dimensional representations for messages with an attentional RNN. Using the message representation as guidance, the POI memory network exploits POI metadata to bridge semantic space and location space.

enables us to measure how much each word contributes to the location aspect of the entire sentence.

Attention is commonly used in sequence-to-sequence networks, as a form of attending to previous encoder state while generating a new sequence[2, 16]. This has been extended to a general form of attention where an alignment score is computed between an external query vector Q and the sequence states $\{h_1, h_2, \cdots, h_n\}$. Then, the retrieved result is a weighted sum over the hidden states. Compared to directly matching h_t and Q, 'attention' can be seen as soft retrieval. A, W_1, W_2 are weight matrices and q is the final representation of the sentence.

$$\begin{aligned} a_t &= A^T \tanh(W_1 Q + W_2 h_t) \\ a'_t &= \frac{\exp(a_t)}{\sum_i \exp(a_i)} \\ q &= \sum_t a'_t h_t \end{aligned} \tag{4.2}$$

When an external query vector Q is not available, we can still obtain specific aspects from the sentence using self-attention [13]. In our case, the message attention layer acts as a 'location extractor'. The attention parameter A acts as an anchor for 'location' related words in semantic space. Formally, the attention scores for words are computed as follows to generate the final message representation q:

$$\begin{aligned} a_t &= A^T \tanh(W_a h_t) \\ a'_t &= \frac{\exp(a_t)}{\sum_i \exp(a_i)} \\ q &= \sum_t a'_t h_t \end{aligned} \tag{4.3}$$

In the case studies we will showcase several examples of how the attention layer successfully identifies location-related words from social media messages.

4.3 The Key-Value Memory Network The memory network leverages existing information sources to guide the mapping from text to geo-locations.

A straightforward approach to do so would be to directly learn a function that takes text and translates it into to locations. However, this approach suffers from two major drawbacks. First of all, training an accurate mapping requires a large amount of training data that covers all areas and all possible text references. Second, such a black-box model provides little insight in the internal working process of the mapping. It is hard to see what drives the model to come to such a prediction.

In comparison, our strategy takes advantage of existing POI metadata with a key-value memory network. Using POI information as an auxiliary information source bootstraps the mapping between text to location. The memory network then learns low-dimensional representations of the textual aspect of POIs that shares a common embedding space with the message representation. In such a way, the model is able to match POIs that have close semantics with the given message to determine the probability of the message originating from that location. Our memory network also introduces another attention layer over POIs, providing interpretability in the prediction process.

4.3.1 Key-Value Embeddings Each entry in the memory network is a single POI $p = (d, l) \in P$, and consists of two aspects: text and location. The text fields in its metadata (name, category etc.) are used to initialize the key k and its location is used as the

corresponding value v.

Each word in the text field is embedded using a shared embedding layer with the message encoder, encouraging the alignment of the two representations, as shown below:

$$d_i = \{u'_1, u'_2, \cdots, u'_n\}$$
$$k_i = \frac{1}{n}\sum_j \Phi(u'_j) \tag{4.4}$$

We directly used average pooling over the embeddings as the POI metadata in our dataset was relatively short. When long textual descriptions are available, it would be suitable to use a RNN similar to that in the message encoder to generate the key embedding.

For the location representation, we have to bear two considerations in mind: the message may have multiple possible matching candidate POIs and the spatial proximity between candidates affect the outcome of prediction. Thus we retreat from directly using coordinates, since multiple candidates will pull the prediction in different directions and the resulting prediction will lie in the middle. We also choose not to use one-hot vectors of manually divided grids as the proximity relationship between POIs are no longer preserved in this representation. In our network, each of the locations v_i are represented by a bi-variate Gaussian distribution over space, centered at the POI coordinates l.

$$v_i = \mathcal{N}(l_i, \Sigma_i) \tag{4.5}$$

It is then natural to reflect our belief about the possible origin of the message using the weighted sum or mixture of memory values.

4.3.2 The POI Attention In the POI key-value memory network, we use another attention layer to selectively focus on the POIs that are relevant to the given message. Instead of selecting the top-k candidates, we softly attend over the entire memory to preserve the end-to-end property of the network.

The external query vector here is the message representation q and the attention weights are computed by aligning q to the keys of the memory network $\{k_1, k_2, \cdots, k_n\}$. This alignment weight is then used to determine the relative weight of the corresponding memory value.

$$a_i = q^T W_m k_i$$
$$\pi_i = \frac{\exp(a_i)}{\sum_j \exp(a_j)} \tag{4.6}$$

The attention score is used as the mixture weight of the Gaussian components, which is then combined to output a Gaussian mixture distribution over the geographic space.

For a given social media message m, the distribution of the source location l is estimated as:

$$Pr(\hat{l}|m) = \sum_{i=1}^{p_n} \pi_i \mathcal{N}(\hat{l}|l_i, \Sigma_i) = \sum_{i=1}^{p_n} \{\pi_i \frac{\exp(-\frac{1}{2}(x-l_i)^T \Sigma_i (x-l_i))}{2\pi\sqrt{\Sigma_i}}\} \tag{4.7}$$

When training the entire network, the loss function is the negative loglikelihood of the Gaussian mixture model over all training examples.

$$\mathcal{L} = -\sum_{(m,l)\in\mathcal{C}} \log\{\sum_{i=1}^{p_n} \pi_i \mathcal{N}(l|l_i, \Sigma_i)\} \tag{4.8}$$

5 Experiments

5.1 Experimental Setup

5.1.1 Dataset We use a real-life geo-tagged tweet dataset collected from New York users. The geo-tagged tweets are collected through Twitter's public API during the period of Aug. 1st - Nov. 30th, 2014, summing to a total of 1.9 million geo-tagged tweets. Each tweet consists of a text message, its original location in coordinates and the timestamp.

We also use a POI dataset collected from Foursquare, which includes a total of 266,291 POI listings. In our experiments we only use the most popular 4000 POIs. We experimented with using more POIs but the improvement was marginal. Each POI is characterized by name, category, and GPS location.

5.1.2 Baselines

- **LR**[25] is a logistic regression model that uses bag-of-words unigrams as features. [1]
- **LGTA** [38] is a geographical topic model that discovers spatially coherent topics from geo-tagged text.
- **CrossMap** [40] is a state-of-the-art approach for spatiotemporal activity modeling. Geodesic grids and words are used to construct a bi-partite network, which is then embedded in low-dimensional space. The grid with the smallest cosine distance to the word embeddings is the predicted location.
- **MDN-Shared** [24] represents the message as a bag-of-words and uses the mixture density network[3].
- **AttnReg** passes the RNN-encoded message through a feed forward network to predict the coordinates.

[1] Grids of 100m × 100m are used.

Method	Acc@1km	Acc@5km	Mean/m	Median/m
LR	**0.1722**	0.3865	9370.12	7671.37
LGTA	0.0249	0.2118	12034.30	13596.33
CrossMap	0.1375	0.2449	13114.57	11688.08
MDN-Shared	0.0424	0.3965	7528.63	6353.75
AttnReg	0.0139	0.2519	8431.57	7976.62
GeoAttn	0.1187	**0.5218**	**6359.14**	**4616.79**
- Memory	0.0599	0.4248	7222.69	6067.92
- Attn	0.0323	0.3067	9015.97	8734.26

Table 1: Quantitative comparison of baselines and model ablations.

To evaluate the effectiveness of the different modules in GeoAttn, we also include ablations of our model for comparison.

- **GeoAttn- Attn** removes the attention layer and represents the message by the last hidden layer of the recurrent neural network.
- **GeoAttn- Mem** is a combination of our model and MDN-Shared. The attention generated message representation is used as the input of the mixture density network.

5.1.3 Implementation We sort the dataset by the timestamp and hold out the most recent 20% of the data for testing for all models. Our word embeddings are trained using the GloVe algorithm on the training set of Twitter messages to avoid data leakage. Then within the remaining 80%, we take another 20% as the validation set to tune the parameters of our model and use the rest as the training set.We use an existing TweetNLP[21] tool to pre-process tweet text which includes tokenizer and part-of-speech tagger. After tokenization, the text is normalized using a dictionary[8]. Our code is publicly available [2] and additional details are available in the supplementary materials.

5.2 Quantitative Evaluation We follow the mainstream literature [4, 9, 34, 35] in geolocation prediction and use three metrics all originated from prediction error distance: **accuracy**, **mean distance** and **median distance**. The distance is computed with the haversine formula, which yields the great-circle distance between two points on a sphere. In our localization setting, we set two thresholds for correct prediction: 1km and 5km. For baseline methods that output a single grid label, the distance is computed from the center of the true label grid to the predicted grid.

Table 1 shows the performance of different methods in terms of the four metrics. Figure 2 looks into the more detailed distribution of prediction error distance

[2] https://github.com/raspberryice/geo-attn

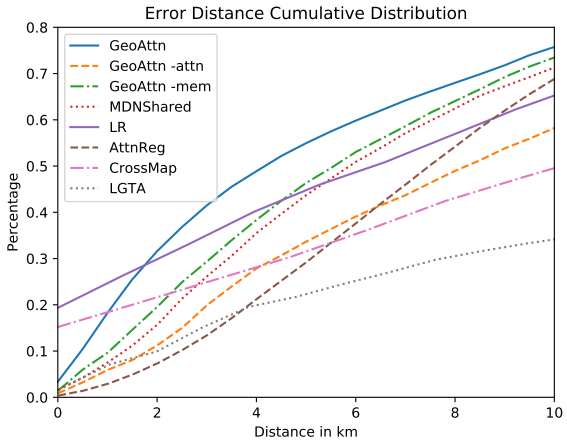

Figure 2: Cumulative distribution of error distance under 10km.

within 10km. Jointly considering the two, we have the following observations:

- LR and CrossMap both treat the localization problem as multi-class classification while all the other methods produce predictions in continuous space. This is reflected in the cumulative distribution curve for LR and CrossMap by not crossing the origin. Given a reasonable grid division heuristic, classification-based methods can achieve good grid-level accuracy. However, since they tie words to discrete grids, they cannot generalize the knowledge to predict labels that do not appear in the training data. As a result, their mean error distance is significantly larger than our purposed model that produces a probability distribution and their accuracy drops when the threshold grows larger.
- As a representative of the topic modeling approach, LGTA does not have strong predictive power. In LGTA, areas are defined as a distribution over words. However, most of the frequent words are location invariant, and the discriminative words are of relatively low frequency, such as POI names. In contrast, GeoAttn utilizes the attention mechanism to recognize location-indicative words.
- The performance gap between the GeoAttn-Attn /GeoAttn and also MDN-Shared/GeoAttn-Mem demonstrates the power of the message attention layer in accurately 'extracting' the location related information from the text message. The attention layer ignores irrelevant content and put weights on only the location-indicative words.
- AttnReg and GeoAttn-Mem differ in the location representation. Feed-forward neural network layers act as function approximators as AttnReg

General Category	Category	Percentage	Total
Predictable	POI Mention	29%	57%
	Region Mention	20%	
	Semantic Clue	18%	
	Event Mention	4%	
Unpredictable	Irrelevant	34%	36%
	Mismatch	2%	
Model Restriction	Information Lost	4%	7%
	Personal	3%	

Table 2: Categories of prediction difficulty

directly learns a mapping from semantic space to coordinates while GeoAttn-Mem learns the mapping from semantic space to Gaussian mixture parameters. Comparing GeoAttn-Mem to AttnReg shows the superiority of predicting a probability distribution instead of a single point in space, since it is possible to have several candidate locations that pull the single output in different directions.

- Comparing our purposed model and GeoAttn-Mem,GeoAttn successfully exploits existing POI metadata to help bridge the gap between semantic space and location, resulting in an 4.4-9.7% boost in accuracy and a 11.1% reduction in mean error distance.

5.3 Data Analysis and Case Studies When it comes to location inference, the first question to ask is "is the location predicable at all?". To answer this question, we randomly selected 100 tweets from the test data and labelled them according to the presence of location-related clues as shown in Table 2. Cases are examined for each category[3] in Figure 3.

5.3.1 Message of exact mention We first examine the most straightforward case: when the POI name is directly mentioned in the message. For Figure 3a *"Happy birthday@Mamajuana Cafe NYC"* the message attention also manages to identify the word *Mamajuana* as location indicative. In the memory network, this name has multiple matches and this is reflected by two peaks in the output distribution corresponding to the two real locations of *Mamajuana Cafe*.

5.3.2 Messages of semantic similarity For this case, the POI name is not directly mentioned but the semantics of the message give hints about the location. The tweet shown in Figure 3b *"Come enjoy a glass of Nero or Pinot Grigio for happy hour! ... #winelover #wineoclock #wine"* mentions wine names and attaches many hashtags related to wine , suggesting that the tweet is posted from a bar. The message attention captures the phrase *a glass of Nero* and matches it to many bar and restaurant locations in Manhattan. In such cases nailing the exact location of the message poses difficulty but we can narrow down the range to make a good estimate. This example shows that our model goes beyond simple symbolic matching with gazettes and leverages semantic similarity.

5.3.3 Messages of region mention In some of the tweets, the location is referred to in a more coarse granularity than exact POIs. For example, the tweet *"#Nursing #Job in #Montclair, ...* in Figure 3c points us to the town Montclair in New Jersey. Our model recognizes the hashtag #Montclair as the location and the POIs that are in the neighbor are assigned with high weights, contributing to the final prediction.

5.3.4 Location Mismatch Our model imposes a strong assumption that the location-related words in the text message are indicators of the origin of the message. However, this assumption does not always hold. In the tweet *"Can't believe I got a ticket in Irvington last night"*, the user is reminiscing in joy from yesterday. Although *Irvington* is correctly recognized as location-indicative, our model fails to account that this is not the current location of the user any more.

5.3.5 Irrelevant Messages The fact that the user chooses to add a geotag to his/her message does not necessary mean that the message itself is closely related to a particular location. The example in Figure 3e *"I think the idea of the gov increasing the alcohol consumption age to 25 ..."* is expressing the user's opinion on the policy. The distribution shows that there is no particular area that matches both words and the prediction confidence is low.

5.3.6 Personal Locations Some of the location mentions, such as *this office*, *home*, *our hotel* are too vague to use for prediction on a single message level. In Figure 3f the tweet *"My girl finally got a dining table for her apartment ..."*, the words *dining* and *apartment* are attached with the largest weight, but it is impossible to know the exact location.

6 Conclusions and Future Work

We have studied the problem of localization for social media messages. To handle the noisy nature of social media messages and take advantage of existing POI metadata, we propose an attentional memory network model named GeoAttn. The entire framework is end-to-end trainable and offers interpretable predictions via attention scores. Our experiments on a million-scale tweet dataset shows that GeoAttn outperforms state-of-the-art methods for localization of social media messages, and meanwhile provides meaningful explanations for its predictions.

In the data analysis section we discover that one-

[3]Standards for each category are explained in the supplementary materials.

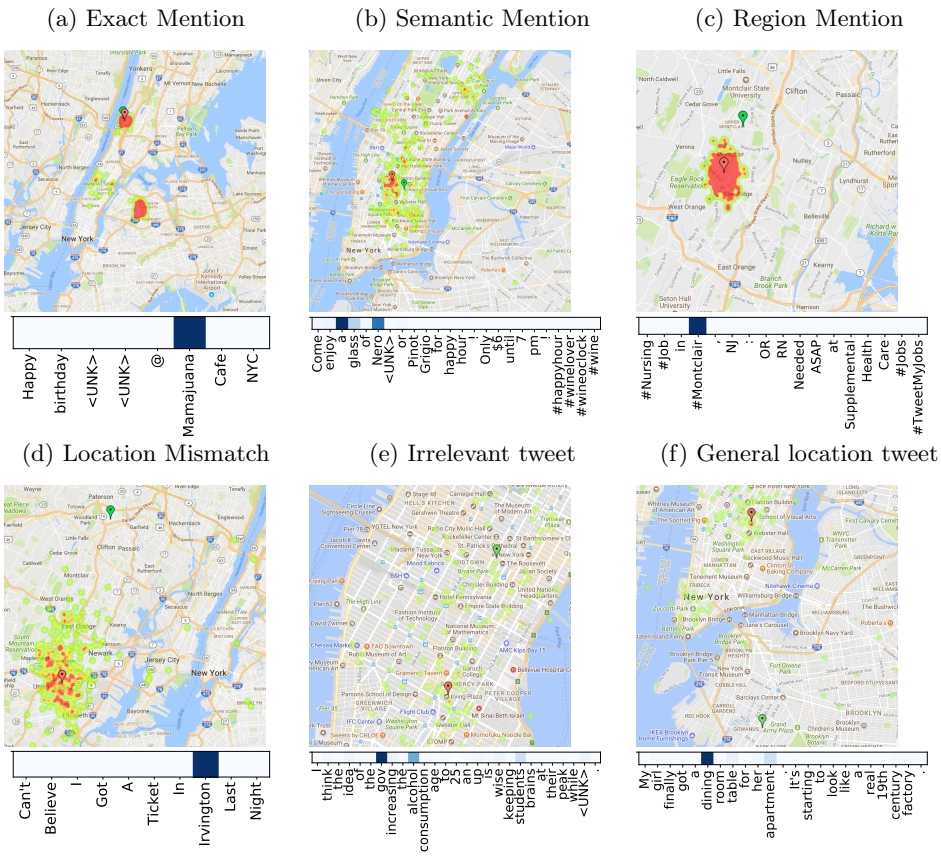

Figure 3: Case study: we show the attention weights applied to each word in twitter messages and the output probability distribution. The green label shows the true location of the tweet and the red label is the predicted location.

third of the messages are not location indicative. The presence of such noisy data degrades the performance of predictive models since it presents misleading signals. A potential direction for future work is to identify such messages are unpredictable. Another possible improvement is to take user history into account and make personalized predictions.

Acknowledgements. Research was sponsored in part by U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), DARPA under Agreement No. W911NF-17-C-0099, National Science Foundation IIS 16-18481, IIS 17-04532, and IIS-17-41317, DTRA HDTRA11810026, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov). Any opinions, findings, and conclusions or recommendations expressed in this document are those of the author(s) and should not be interpreted as the views of any U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

[1] A. Ahmed, L. Hong, and A. J. Smola, *Hierarchical geographical modeling of user locations from social media posts*, in WWW, 2013, pp. 25–36.

[2] D. Bahdanau, K. Cho, and Y. Bengio, *Neural machine translation by jointly learning to align and translate*, ICLR, (2015).

[3] C. M. Bishop, *Mixture density networks*, (1994).

[4] Z. Cheng, J. Caverlee, and K. Lee, *You are where you tweet: a content-based approach to geo-locating twitter users*, in CIKM, 2010, pp. 759–768.

[5] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, *Learning phrase representations using rnn encoder-decoder for statistical machine translation*, in EMNLP, 2014, pp. 1724–1734.

[6] R. Compton, D. Jurgens, and D. Allen, *Geotagging one hundred million twitter accounts with total variation minimization*, Big Data, (2014), pp. 393–401.

[7] J. Eisenstein, B. T. O'Connor, N. A. Smith, and E. P. Xing, *A latent variable model for geographic lexical variation*, in EMNLP, 2010, pp. 1277–1287.

[8] B. Han, P. Cook, and T. Baldwin, *Automatically constructing a normalisation dictionary for microblogs*, in EMNLP-CoNLL, 2012, pp. 421–432.

[9] ———, *Text-based twitter user geolocation prediction*, J. Artif. Intell. Res., 49 (2014), pp. 451–500.

[10] J. He, W. Shen, P. Divakaruni, L. Wynter, and R. Lawrence, *Improving traffic prediction with tweet semantics*, in IJCAI, 2013, pp. 1387–1393.

[11] D. Jurgens, *That's what friends are for: Inferring location in online social media platforms based on social relationships*, in ICWSM, 2013.

[12] M. D. Lieberman, H. Samet, and J. Sankaranarayanan, *Geotagging with local lexicons to build indexes for textually-specified spatial data*, ICDE, (2010), pp. 201–212.

[13] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, *A structured self-attentive sentence embedding*, ICLR, (2017).

[14] J. Liu and D. Inkpen, *Estimating user location in social media with stacked denoising auto-encoders*, in VS@HLT-NAACL, 2015, pp. 201–210.

[15] Y. Liu, C. Liu, X. Lu, M. Teng, H. Zhu, and H. Xiong, *Point-of-interest demand modeling with human mobility patterns*, in KDD, 2017, pp. 947–955.

[16] T. Luong, H. Pham, and C. D. Manning, *Effective approaches to attention-based neural machine translation*, in EMNLP, 2015.

[17] A. H. Miller, A. Fisch, J. Dodge, A.-H. Karimi, A. Bordes, and J. Weston, *Key-value memory networks for directly reading documents*, in EMNLP, 2016, pp. 1400–1409.

[18] Y. Miura, M. Taniguchi, T. Taniguchi, and T. Ohkuma, *Unifying text, metadata, and user network representations with a neural network for geolocation prediction*, in ACL, 2017, pp. 1260–1272.

[19] L. Moncla, W. Renteria-Agualimpia, J. Nogueras-Iso, and M. Gaio, *Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus*, in SIGSPATIAL/GIS, 2014, pp. 183–192.

[20] S. A. Moorhead, D. E. Hazlett, L. Harrison, J. K. Carroll, A. Irwin, and C. Hoving, *A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication*, Journal of medical Internet research, 15 (2013).

[21] O. Owoputi, B. T. 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, 2013, pp. 380–390.

[22] J. Pennington, R. Socher, and C. D. Manning, *Glove: Global vectors for word representation*, in EMNLP, 2014, pp. 1532–1543.

[23] Y. Qian, J. Tang, Z. Yang, B. Huang, W. Wei, and K. M. Carley, *A probabilistic framework for location inference from social media*, ArXiv, abs/1702.07281 (2017).

[24] A. Rahimi, T. Baldwin, and T. Cohn, *Continuous representation of location for geolocation and lexical dialectology using mixture density networks*, in EMNLP, 2017, pp. 167–176.

[25] A. Rahimi, T. Cohn, and T. Baldwin, *Twitter user geolocation using a unified text and network prediction model*, in ACL, 2015, pp. 630–636.

[26] ———, *A neural model for user geolocation and lexical dialectology*, in ACL, 2017, pp. 209–216.

[27] S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge, *Supervised text-based geolocation using language models on an adaptive grid*, in EMNLP-CoNLL, 2012, pp. 1500–1510.

[28] K. Ryoo and S. Moon, *Inferring twitter user locations with 10 km accuracy*, in WWW, 2014, pp. 643–648.

[29] A. Sadilek, H. A. Kautz, and J. P. Bigham, *Finding your friends and following them to where you are*, in WSDM, 2012, pp. 723–732.

[30] R. Steele, *Social media, mobile devices and sensors: categorizing new techniques for health communication*, in Sensing Technology (ICST), 2011, pp. 187–192.

[31] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, *End-to-end memory networks*, in NIPS, 2015, pp. 2440–2448.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention is all you need*, in NIPS, 2017, pp. 6000–6010.

[33] J. Weston, S. Chopra, and A. Bordes, *Memory networks*, ICLR, (2015).

[34] B. Wing and J. Baldridge, *Simple supervised document geolocation with geodesic grids*, in ACL, 2011, pp. 955–964.

[35] ———, *Hierarchical discriminative classification for text-based geolocation*, in EMNLP, 2014, pp. 336–348.

[36] H. Xu and K. Saenko, *Ask, attend and answer: Exploring question-guided spatial attention for visual question answering*, in ECCV, 2016, pp. 451–466.

[37] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, *Show, attend and tell: Neural image caption generation with visual attention*, in ICML, 2015, pp. 2048–2057.

[38] Z. Yin, L. Cao, J. Han, C. Zhai, and T. S. Huang, *Geographical topic discovery and comparison*, in WWW, 2011, pp. 247–256.

[39] C. Zhang, L. Liu, D. Lei, Q. Yuan, H. Zhuang, T. Hanratty, and J. Han, *Triovecevent: Embedding-based online local event detection in geo-tagged tweet streams*, in KDD, 2017, pp. 595–604.

[40] C. Zhang, K. Zhang, Q. Yuan, H. Peng, Y. Zheng, T. Hanratty, S. Wang, and J. Han, *Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning*, in WWW, 2017, pp. 361–370.

[41] S. Zhao, T. Zhao, I. King, and M. R. Lyu, *Geo-teaser: Geo-temporal sequential embedding rank for point-of-interest recommendation*, in WWW, 2017, pp. 153–162.