GeoAttn: Fine-Grained Localization of Social Media Messages via Attentional Memory Network
Supplementary Materials

Sha Li, Chao Zhang, Dongming Lei, Ji Li, Jiawei Han
Dept. of Computer Science, University of Illinois at Urbana-Champaign
{shal2, czhang82, dlei5, jili3, hanj}@illinois.edu

1 Data Analysis Details

We use the following categories for classifying the predictability of social network messages:

- **POI Mention** A venue name is mentioned in the message. This is the strongest type of clue for localization and thus the easiest case.
- **Region Mention** A coarse grain neighborhood or community name is mentioned in the message.
- **Semantic Clue** Some words in the message give clues about the location of origin. The location can be inferred with high uncertainty.
- **Event Mention** An event is mentioned in the message. As an event is temporally, there may be few historical messages to infer the location of the event.
- **Irrelevant** The topic that the message discusses does not contain any clues of the location.
- **Location Mismatch** The location inferred from the message is not the actual geotag of the message. This may be because the message is referring to an event earlier or later in time.
- **Information Lost** The critical location-related information in the message is lost due to earlier preprocessing steps, including tokenization, vocabulary building etc.
- **Personal** The location-related terms are too broad to infer a particular geo-location but with the help of more user-specific information, the location can be inferred.

The first four categories are predictable given our model. The irrelevant cases and the location mismatch cases violate the basic assumption of our model that is the text information is related to the location of the message. Thus these cases are impossible for us to predict. The information lost case can be reduced with better preprocessing, and the personal location case can be alleviated by considering the user's personal history.

2 Implementation Details

We use an existing TweetNLP(2) tool to pre-process tweet text which includes tokenizer and part-of-speech tagger. After tokenization, the text is normalized using a dictionary(1).

When training GloVe vectors, we set the vocabulary cutoff $\tau = 10$, amounting to a total of 59023 words. During training, the window size is set to 15. The resulting vectors are of 200 dimensions.

Within the neural network, the hidden unit size for GRU is 128. Dropout is performed after embedding layers and throughout the GRU with a dropout rate of 0.4. For the memory network and MDN, the number of components and the number of memory slots are both set to 4000 for fair comparison.

We implement our model with the Pytorch framework. Our network is optimized using stochastic gradient descent with the optimizer Adamax.

Early termination of training was determined by the validation performance to alleviate overfitting.

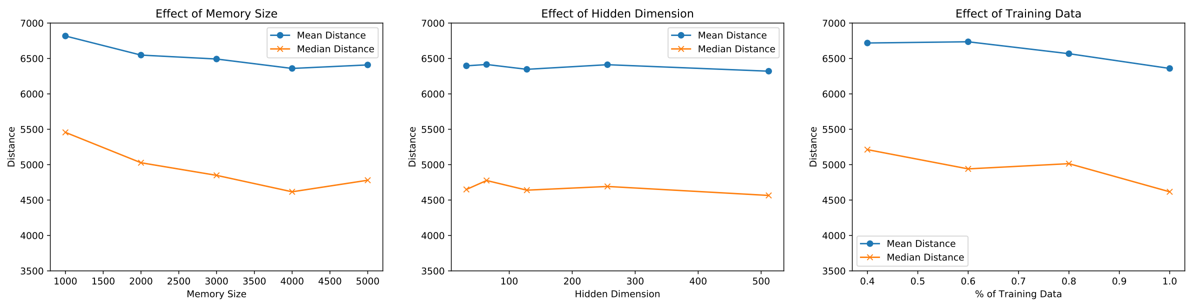
3 Parameter Study

Using our GeoAttn model as a basis, we tune the major hyper-parameters of our model and report the corresponding error distance in Figure 1.

When tuning the memory size K, we sort the POIs by their popularity and take the top-K POIs to initialize the memory entries. Both mean distance and median distance plateau as the memory size reaches 4000. Therefore, we set 4000 as our optimal memory size and also use the same number of components in MDN-Shared for fair comparison.

The effect of the hidden layer dimension in the RNN is not as significant. We also experimented with using stacked GRUs with 2 or 3 layers instead of just a single layer but the accuracy difference was below 1% and the mean distance improved within 100m.

Regarding the amount of training data, apparently the more the better, but we wish to point out that



(a) Effect of memory size

(b) Effect of hidden layer h

(c) Effect of amount of training data. 100% refers to original training set

Figure 1: Performance of model under different parameters

our model outperforms baselines even when using only 40% of the training data. This is because our model leverages existing POI metadata to bridge the gap between modalities, comparing to other methods that try to learn an accuracy mapping from scratch.

References

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

[2] 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.