GEO-BLEU: Similarity Measure for Geospatial Sequences

Toru Shimizu [1], **Kota Tsubouchi** [1], **Takahiro Yabe** [2]
[1] Yahoo Japan Corporation,
Tokyo, Japan
[2] MIT Media Lab,
Cambridge, MA, USA
{toshimiz, ktsubouc}@yahoo-corp.jp, tyabe@mit.edu

Abstract

In recent geospatial research, the importance of modeling large-scale human mobility data via self-supervised learning is rising, in parallel with progress in natural language processing driven by self-supervised approaches using large-scale corpora. Whereas there are already plenty of feasible approaches applicable to geospatial sequence modeling itself, there seems to be room to improve with regard to evaluation, specifically about how to measure the similarity between generated and reference sequences. In this work, we propose a novel similarity measure, GEO-BLEU, which can be especially useful in the context of geospatial sequence modeling and generation. As the name suggests, this work is based on BLEU, one of the most popular measures used in machine translation research, while introducing spatial proximity to the idea of n-gram. We compare this measure with an established baseline, dynamic time warping, applying it to actual generated geospatial sequences. Using crowdsourced annotated data on the similarity between geospatial sequences collected from over 12,000 cases, we quantitatively and qualitatively show the proposed method's superiority.

Introduction

Geospatial sequence modeling over human mobility trajectories and language modeling in natural language processing (NLP) can be seen analogously, regarding places as words and human mobility trajectories as sentences. On the geospatial side, the main workhorse is next place prediction (NPP), in which a model predicts the place a person moves to at the next time step on the basis of the past trajectory and other features, and repeating NPP while reusing predicted places as context directly leads to geospatial sequence generation. Also, this approach can naturally extend to sequence-to-sequence or translation, assuming a model generates a trajectory using another corresponding trajectory, e.g., one in a past period, as context. The importance of this kind of self-supervised approach is surging in geospatial research, and many modeling methods known in NLP and other related fields are feasibly applicable to geospatial problem settings. Meanwhile, the area of evaluation still seems to be needing further consideration.

- Dynamic time warping (DTW) (Vintsyuk 1968; Sakoe and Chiba 1978) has long been known as a way to evaluate the distance of two given sequences, and it has been used in geospatial research as well as in many other fields. An essential characteristic of DTW is that it aligns the sequences for measuring entirely, without considering local features shared between them. It is suitable to treat entirely aligned sequences, but not so when treating those not aligned.
- BLEU (Papineni et al. 2002) is one of the most popular measures for similarity used in NLP, especially in machine translation. BLEU uses local features of given sequences, word n-grams, and is suitable to treat not completely aligned sequences. Regarding places in sequences as words and their contiguous combinations as geospatial n-grams, we can apply this to evaluate the similarity of geospatial trajectories on the basis of local features. However, it has another disadvantage; the geospatial n-grams need to be exactly the same to be counted as "matched", and very close but slightly displaced ones do not contribute to the outcome. In other words, spatial proximity, which is potentially an important property for similarity, is not taken into account when using BLEU.

In this work, we propose a novel alternative, GEO-BLEU, extending BLEU to incorporate the idea of geospatial proximity into its core concept. To evaluate the measure's performance, we use a translation problem of human mobility trajectories as a plausible test case for similarity and distance measures; we collected trajectories consisting of smartphone locations and modeled them in a sequence-to-sequence manner. The modeling task is to predict peoples' daily trajectories under the self-restraint of COVID-19 on the basis of trajectories in the pre-COVID-19 period, which ultimately leads to understanding behavioral implications of COVID-19 and contributing to the field of urban dynamics. We apply our proposed measure, GEO-BLEU, and other two baselines, BLEU and DTW, to sequences generated by this translation model, comparing generated sequences and actual sequences. After that, we compare these scores with crowdsourced annotated data to quantify how consistent the measures and human intuition are, showing the proposed method's superiority.

Existing and Proposed Measures

In this section, we first explain DTW and BLEU and then describe our proposed measure GEO-BLEU. Also, using a toy problem, we demonstrate a notable characteristic of the proposed method.

Existing Measures

Dynamic Time Warping. Dynamic time warping (DTW) (Vintsyuk 1968; Sakoe and Chiba 1978) is a distance-like measure for comparing the similarity between two sequences which was first developed in speech recognition but then has been used in various fields including geospatial research. The method involves finding the optimal alignment between two sequence $X = (x_1, x_2, \dots, x_M)$ and $Y = (y_1, y_2, \dots, y_N)$. One possible way of alignment is represented as a sequence of pairs between elements in X and those in Y: $P = ((x_{i_1}, y_{j_1}), (x_{i_2}, y_{j_2}), \dots, (x_{i_l}, y_{j_l}), \dots, (x_{i_L}, y_{j_L}))$ where $i_l \in [1 : M], j_l \in [1 : N]$ and $L = \max(M, N)$. Also, there are three conditions for P to be valid alignment:

- the boundary condition $(i_1, j_1) = (1, 1)$ and $(i_L, j_L) = (M, N)$, which requires the start of X and Y and the end of them must be matched respectively,
- monotonicity condition $i_l \leq i_{l+1}$ and $j_l \leq j_{l+1}$ for $l \in [1 : L - 1]$, which preserves the time-ordering of elements, and
- step size condition $(i_{l+1} - i_l, j_{l+1} - j_l) \in \{(1, 1), (1, 0), (0, 1)\}$.

The cost for such an alignment P is calculated as the sum of the pairwise distance $d(x_{i_l}, y_{j_l})$:

$$cost(P) = \sum_{l=1}^{L} d(x_{i_l}, y_{j_l}) \tag{1}$$

where $d(\cdot, \cdot)$ is usually the Euclidean distance between two places. Using this, we can represent DTW as the minimum cost given by the optimal P:

$$DTW = \min_{P} cost(P). \tag{2}$$

As for the actual procedure of optimization, we followed a technical report (Senin 2008).

BLEU. BLEU (Papineni et al. 2002) is a measure being heavily used for evaluating machine translation systems for quantifying how close generated candidates are to reference human translations. BLEU uses word n-grams as the unit of comparison and considers the ratio of n-grams matched between the generated and reference sentences to all the n-grams in the generated candidates for a given n. The ratio, which is called modified precision p_n, is obtained as follows

$$p_n = \frac{\sum_{S \in C} \sum_{n\text{-}gram \in S} \mathrm{Count}_{\mathrm{matched}}(n\text{-}gram)}{\sum_{S \in C} \sum_{n\text{-}gram' \in S} \mathrm{Count}(n\text{-}gram')} \tag{3}$$

where C is the candidate corpus, and S is each of the candidate sentences in it. Actually, p_n tends to become large when the candidates are too short. To correct this unintended effect, BLEU uses a factor called the brevity penalty BP, which is given by

$$BP = \begin{cases} 1, & \text{if } c > r \\ e^{1-r/c}, & \text{if } c \leq r \end{cases} \tag{4}$$

where c is the sum of the candidates' lengths, and r is that of the references. Taking the weighted geometric average of the modified precision scores for $n \in \{1, \dots, N\}$ while applying BP, resultant BLEU score B is defined as

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{5}$$

where w_n is the positive weight summing up to 1. The original work of BLEU uses $N = 4$ and $w_n = \frac{1}{N}$ for $n \in \{1, \dots, 4\}$, and we follow the settings in the current study. It should be noted that BLEU is for evaluating candidate and reference sentences of the whole corpus and not for evaluating a single candidate sentence.

GEO-BLEU

Our proposed measure GEO-BLEU is based on BLEU but intended to be an alternative to DTW, which means it measures a distance or similarity of a given pair of sequences. At the same time, it borrows the concept of n-gram from NLP, relaxing the matching condition so that the score reflects the proximity of a given pair of n-grams.

As the first step, we introduce the geospatial revision of n-gram as a chunk of locations $(q_1, \dots, q_n)$ where each location q_k is represented as a point in two-dimensional space. In addition, we define the similarity score s of a pair of n-grams $g_v = (v_1, \dots, v_n)$ and $g_w = (w_1, \dots, w_n)$ on the basis of proximity as follows

$$s(g_v, g_w) = \prod_{k=1}^{n} \exp(-\beta\, d(v_k, w_k)) \tag{6}$$

where $d(\cdot, \cdot)$ is the Euclidean distance between two locations, and β is a coefficient for adjusting the scale. In this manner, the similarity between n-grams is evaluated to become one when two n-grams are exactly matched. Also, the far two n-grams go away, the closer the value asymptotically comes to zero.

Next, we consider the way to match n-grams in the candidate sequence and those in reference. In BLEU, the matching is conducted by the function $Count_{\mathrm{matched}}(n\text{-}gram)$ in Equation 3; it gives one if the same n-gram remains "unused" in the reference sentences, eliminating that "used" n-gram instance from the pool for subsequent matching, and otherwise gives zero. For GEO-BLEU, which incorporates the concept of proximity, we let an n-gram on the candidate side form a pair with the closest unused n-gram remaining on the reference side, prohibiting n-grams on the reference side from being reused as in the BLEU's original matching rule. We greedily optimize the set of such pairs so that the sum of the similarity scores comes close to the maximum value. Denoting the optimized set of pairs as $P = \{(g_{c_1}, g_{r_1}), \dots, (g_{c_L}, g_{r_L})\}$ where L is the shorter of the candidate's and reference's lengths, g_{c_k} is an n-gram of the candidate sequence, and g_{r_k} is that of the reference sequence, we define our n-gram-based similarity q_n for a pair

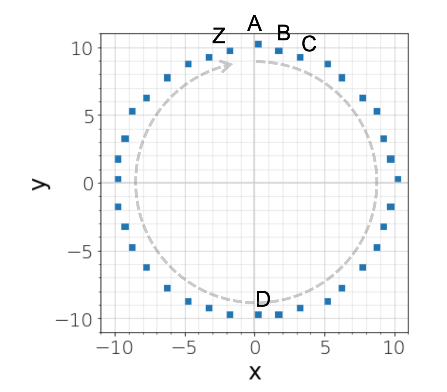

Figure 1: A sample sequence consisting of 36 grid cells placed over a circle of 10 km radius.

of a candidate sequence S and its reference sequence as

$$q_n = \frac{\sum_{(g_c, g_r) \in P} s(g_c, g_r)}{\sum_{n\text{-}gram \in S} Count(n\text{-}gram)}. \tag{7}$$

Taking the weighted geometric mean for a range of n in the same manner as Equation 5 and introducing the brevity penalty BP as in Equation 4, the proposed similarity measure GEO-BLEU is given as

$$GEO\text{-}BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log q_n\right). \tag{8}$$

In our experiments, we use $\beta = 1$, $N = 3$, and $w_n = \frac{1}{N}$ for $n \in \{1, 2, 3\}$.

If BLEU is applied to evaluating a single candidate, there can be cases in which the modified precision becomes zero. On the contrary, the modified-precision equivalent of GEO-BLEU always has a non-zero value due to the relaxed matching, and this property makes GEO-BLEU more feasible and suitable for evaluating a single candidate sequence.

Characteristics of GEO-BLEU To illustrate the characteristics of GEO-BLEU and its difference from DTW, we apply the two measures to simple toy sequences in two-dimensional space and compare the results. As shown in Figure 1, we consider 36 grid cells with sides of 0.5 km placed over a circle of 10 km radius at almost regular intervals. Our original sample sequence starts from cell A, goes clockwise through B, C, and the following, and ends at Z as shown as the dashed arc arrow. Then, by moving the start and end points clockwise and one step at a time, i.e., by shifting the phase forward, we can generate variations such as one going clockwise from B to A, another from C to B, and so on for evaluating the similarity with or distance from the original. Here, it is crucial that whether they are similar or different depends on the evaluations' purpose and point of view, and there is no definite criterion in that regard. Considering the original sequence and another with the opposite phase starting from D, they are completely different when aligned wholly. In this view, the distance between the first cells of the sequences is 20 km, the maximum possible number in this setting, and it does not change in the following aligned pairs, such as one between the second cells of the two sequences. On the other hand, those two sequences can be seen as almost identical when concerned with the local features, as they share almost all the cells and chunks except for those around the start and end. Among these conflicting points of view, GEO-BLEU is for comparing sequences on the basis of local features as in the latter example, while DTW views two sequences wholly aligned as in the former.

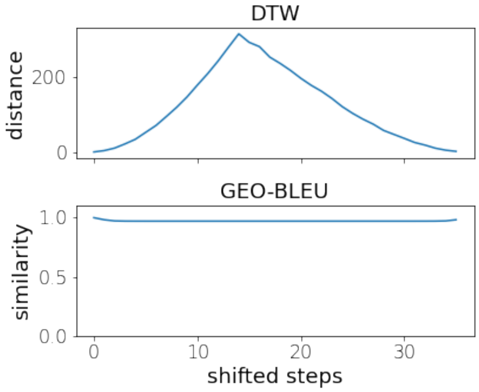

Figure 2: The scores of DTW and GEO-BLEU depending on the extent of the phase shift between the original and its shifted, derived sequence.

Figure 2 shows the actual distance calculated by DTW and similarity by GEO-BLEU between the original and shifted sequences where the x-axis denotes the number of the shifted steps. The results are contrasting; the value of DTW is significantly changing depending on x, while that of GEO-BLEU is staying around the maximum possible value as two sequences are always similar considering their local features. As illustrated, GEO-BLEU is a measure for comparing sequences on the basis of their partial or local features and without aligning them wholly.

Experiments

Human Mobility Trajectory Dataset

For this study, a web service company, Yahoo! JAPAN, provided us with smartphone GPS records of their users, which had originally been collected for their services. The users have agreed to provide their location information for research purposes, and the data are anonymized so that individuals cannot be identified and that personal properties such as gender and age are unknown. Each GPS record consists of a user's ID, timestamp, longitude, and latitude.

Using this smartphone GPS data, we extracted records from two consecutive periods, one from Oct. 1st, 2019 to Mar. 31st, 2020 and the other from Apr. 1st to May 25th in 2020. The periods were determined so that the data captures two different modes of the society in Japan, one mode without the influence of COVID-19, which corresponds to the former period, and the other mode under nationwide self-

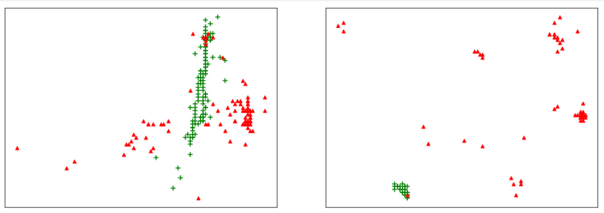

Figure 3: An example of annotation cases for evaluating each measure. The red dots show the steps of a generated trajectory, and the green dots show the steps of its corresponding actual trajectory. In this example, the left side actually belongs to the top-20 pairs and the right side the bottom-20 pairs, and the annotator is asked to distinguish which side comes from the top-20. The easier the cases of a measure are, the better the measure is considered to be.

restraint of activities to prevent the spread of the infection, which corresponds to the latter period. As for deciding the specific dates, we referred to dates of the government's relevant announcements in addition to public stringency index data (Ritchie et al. 2020).

From this set of GPS records, we prepared one million pairs of trajectories representing how the mobility pattern of a person in the pre-COVID-19 period has changed in the mid-COVID-19 period. In this data preparation process, the longitudes and latitudes in the GPS records are aggregated and discretized into 500m-square grid cells on an hourly basis so that a sequence of grid cells corresponds to a trajectory of a person. Then, we allocated 10,000 pairs to the validation set, another 10,000 pairs to the test set, and the rest to the training set. The average length of sequences is 90.5 in the former period and 124.5 in the latter period. Considering that each step stands for an hourly position, each of these sequences usually amounts to a mobility pattern spanning over several days rather than a short trajectory within a day.

Model and Training

We trained a seq2seq model (Sutskever, Vinyals, and Le 2014; Cho et al. 2014) consisting of a pair of two-layer LSTM-RNNs using the dataset so that it can generate a person's trajectory in the mid-COVID-19 period given a trajectory in the pre-COVID-19 period. After the training finished, we took out a model with the lowest validation loss and applied it to sample sequence generation as in the next section. We consider that this approach comes under self-supervised learning as the preparation of the prediction targets does not involve any human judgment or annotations, whereas the model is seq2seq, which is often applied to supervised translation problems.

Evaluation and Results

We generated 500 sample sequences using the trained model and the test set. Then, we scored each sequence with three measures, normalized DTW, GEO-BLEU, and BLEU, comparing the generated trajectory with the actual one in the dataset. As DTW has a dependency on the sequence length as in Equation 1, we normalize the raw scores dividing them by L. Actually, BLEU is a quality measure not of a single sentence but of the entire corpus. However, we dared to apply BLEU for evaluating each sequence here for reference, treating a pair of sequences as a small corpus, to show the difference between the original BLEU and GEO-BLEU.

Method	normalized DTW	**GEO-BLEU**	BLEU
Score	0.550	**0.699**	0.530

Table 1: The average of annotation scores for each measure.

To evaluate how convincing the scores given by a measure are, we sorted the 500 pairs of generated and actual trajectories by each of three measures into descending order of similarity, obtaining three lists of the same entries but in different orders. Then, we took out the top-20 pairs and the bottom-20 pairs from those lists. There are $20 \times 20 = 400$ possible combinations between the top pairs and bottom pairs for each measure, and for each such combination, we asked annotators which pair of generated and actual sequences look more similar, showing the top and bottom pairs side-by-side, as shown in Figure 3. An annotation is given as one of four options: the left is clearly similar, the left is somewhat similar, the right is somewhat similar, and the right is clearly similar. We assign a positive score to a case if the judgment is consistent with the measure: 1.0 if the top-side is judged as clearly similar and 0.5 if it is somewhat similar. If the judgment is inconsistent with the measure, the score becomes -1.0 for "clearly similar" and -0.5 for "somewhat similar" to give a penalty. We presented one case to ten different annotators and collected $400 \times 10 \times 3 = 12,000$ judgments. Table 1 shows the averaged score for three measures, and GEO-BLEU is superior to normalized DTW in this comparison. While the purpose of BLEU is not a single-sequence measurement as in this evaluation, the scores themselves imply that GEO-BLEU derived from it is modified so that it becomes more suitable to the current problem settings.

Related Work

For measuring the quality of a generated sequence compared with a reference sequence, dynamic time warping has primarily been used in geospatial and urban dynamics studies as in various previous works (Bhadane and Shah 2017; Cai et al. 2018). Other types of measures have already been proposed as described in a survey (Su et al. 2020). Still, to the best of our knowledge, this work first applies the concept of "geospatial n-gram" to such evaluation for taking the local features of sequences into account.

Conclusion

We proposed a novel similarity measure of sequences, GEO-BLEU, extending BLEU by incorporating proximity into the core concept and using place n-grams as local features. In a realistic setting about self-supervised geospatial sequence modeling, GEO-BLEU is more consistent with annotators' intuition for similarity than an existing popular measure, DTW.

References

Bhadane, C.; and Shah, K. 2017. Analysis of User Similarity Measures Using GPS Trajectories. In *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, 1–6. IEEE.

Cai, Q.; Liao, L.; Zou, F.; Song, S.; Liu, J.; and Zhang, M. 2018. Trajectory similarity measuring with grid-based DTW. In *International Conference on Smart Vehicular Technology, Transportation, Communication and Applications*, 63–72. Springer.

Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. Doha, Qatar: Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 311–318. USA: Association for Computational Linguistics.

Ritchie, H.; Mathieu, E.; Rodés-Guirao, L.; Appel, C.; Giattino, C.; Ortiz-Ospina, E.; Hasell, J.; Macdonald, B.; Beltekian, D.; and Roser, M. 2020. Coronavirus Pandemic (COVID-19). *Our World in Data*. Retrieved from: https://ourworldindata.org/coronavirus.

Sakoe, H.; and Chiba, S. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1): 43–49.

Senin, P. 2008. Dynamic Time Warping Algorithm Review. Technical report, University of Hawaii at Manoa.

Su, H.; Liu, S.; Zheng, B.; Zhou, X.; and Zheng, K. 2020. A survey of trajectory distance measures and performance evaluation. *The VLDB Journal*, 29(1): 3–32.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, 3104–3112. Cambridge, MA, USA: MIT Press.

Vintsyuk, T. K. 1968. Speech discrimination by dynamic programming. *Cybernetics*, 4: 52–57.