GEO-BLEU: Similarity Measure for Geospatial Sequences

GEO-BLEU: Similarity Measure for Geospatial Sequences

Toru Shimizu
Yahoo Japan Corporation
Tokyo, Japan
toshimiz@yahoo-corp.jp

Kota Tsubouchi
Yahoo Japan Corporation
Tokyo, Japan
ktsubouc@yahoo-corp.jp

Takahiro Yabe
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
Purdue University
West Lafayette, Indiana, USA
tyabe@mit.edu

ABSTRACT

In recent geospatial research, the importance of modeling and generating human mobility trajectories is rising. 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 measuring the similarity between generated and reference trajectories. 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 method, dynamic time warping, applying both measures to simple artificial sequences and examining differences in their characteristics.

CCS CONCEPTS

• **Information systems** → **Spatial-temporal systems**; • **General and reference** → **Metrics**; **Evaluation**.

KEYWORDS

sequence modeling, human trajectory, evaluation

ACM Reference Format:
Toru Shimizu, Kota Tsubouchi, and Takahiro Yabe. 2022. GEO-BLEU: Similarity Measure for Geospatial Sequences. In *The 30th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '22), November 1–4, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3557915.3560951

1 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) [14] 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.

Figure 1: Predicted and actual human mobility trajectories in a relatively short time period, e.g. tens of minutes.

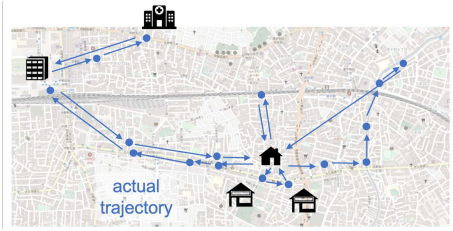

Figure 2: A complex human mobility trajectory in a relatively long time period, e.g. several days.

- Dynamic time warping (DTW) [1, 2] 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 involved sequences for which step-by-step alignment does not make much sense.
- BLEU [9] 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

SIGSPATIAL '22, November 1–4, 2022, Seattle, WA, USA

ACM ISBN 978-1-4503-9529-8/22/11.
https://doi.org/10.1145/3557915.3560951

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.

There can actually be situations where DTW is not suitable. Figure 1 shows predicted and actual trajectories in a relatively short time period, e.g. tens of minutes. In this case, trajectories are simple enough to be aligned in a meaningful way as illustrated by the dotted green lines, and thus DTW is applicable here without problems. On the other hand, when the time period of prediction is relatively long, e.g. several days, trajectories to be predicted will become more involved as illustrated in Figure 2. The trajectory is not a straight line from one place to another anymore but a combination of subtrajectories such as one from home to the office, one from the office to a nearby hospital, and so on. In this problem setting, we can expect that a predicted trajectory shares some motifs or subtrajectories with the actual one locally but not that the whole predicted and actual trajectories can be aligned from the start to the end in a meaningful way, possibly having subtrajectories occurring in a different order.

In this work, we propose a novel alternative, GEO-BLEU, extending BLEU to incorporate the idea of geospatial proximity into its core concept while utilizing local features and not requiring alignment.

2 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.

2.1 Existing Measures

2.1.1 Dynamic Time Warping. Dynamic time warping (DTW) [12, 18] 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 sequences $X = (x_1, x_2, \ldots, x_M)$ and $Y = (y_1, y_2, \ldots, y_N)$. One possible way of alignment is represented as a sequence of pairs between elements in X and those in Y: $P = ((x_{i_1}, y_{j_1}), (x_{i_2}, y_{j_2}), \ldots, (x_{i_l}, y_{j_l}), \ldots, (x_{i_L}, y_{j_L}))$ 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 [15].

2.1.2 BLEU. BLEU [9] 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 $n \in \{1, \ldots, N\}$ 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 $n \in \{1, \ldots, 4\}$, 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. Nevertheless, we borrow the approach of BLEU to devise a measure applicable to a single pair of sentences which can be an alternative to DTW.

2.2 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 $(q_1, \ldots, q_n)$ 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 $g_v = (v_1, \ldots, v_n)$ and $g_w = (w_1, \ldots, w_n)$ 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 $Count_{\text{matched}}(n\text{-}gram)$ 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 $P = \{(g_{c_1}, g_{r_1}), \ldots, (g_{c_L}, g_{r_L})\}$ 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 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.

2.2.1 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 3, 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

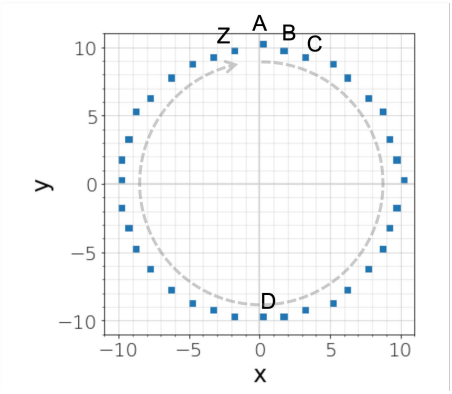

Figure 3: A sample sequence consisting of 36 grid cells placed over a circle of 10 km radius.

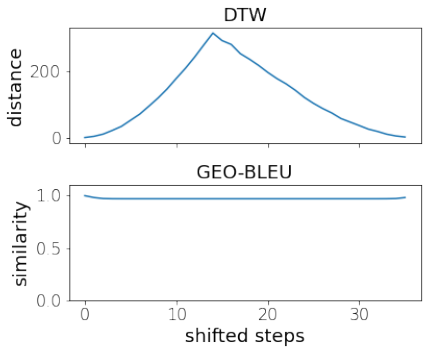

Figure 4: The scores of DTW and GEO-BLEU depending on the extent of the phase shift between the original and its shifted, derived sequence.

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 4 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 subject of comparison is the original sequence itself at $x = 0$, two sequences have the opposite phases at around $x = 18$, and the phase difference becomes very small again at the rightmost point $x = 35$. 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.

3 RELATED WORK

Many studies have proposed methods to measure the similarity of two movement trajectories. First, there are two major types of methods: one considering the entire trajectory and the other considering only a part of the trajectory. The former is called complete match measure and the latter is called partial match measure, as summarized in the following sections. Also, other types of measures have been proposed as described in a survey [16]. Still, to the best of our knowledge, this work is the first to apply the concept of "geospatial *n*-gram" to such evaluation, to take into account the local features of sequences.

3.1 Complete Match Measure

Complete match measure is a method of comparing two trajectories with respect to all the steps.

The most basic method for complete match is the Euclidean distance [17], which calculates the difference in the norm of the trajectories to be compared. It was proposed as a distance measure between time series and had been one of the most widely used distance functions since the 1960s [4, 10]. It is now also used to evaluate movement trajectories. In this case, the trajectories must have the same length.

The most famous algorithm for complete match is Dynamic Time Warping (DTW) [1, 2]. This method has long been used to measure distances in time series data [3, 13], and it is now also used to compare movement trajectories. The algorithm is simple [7], and the lengths of the two trajectories do not have to be the same.

3.2 Partial Match Measure

Partial match measure is a method to measure similarity in only one part of two movement trajectories with a large amount of information. Two well-known methods for partial match are the Longest common subsequence (LCSS) and edit distance on real sequence (EDR) methods.

LCSS measures [5, 11] the length of the sequence common to two trajectories at successive points. For example, two people who were separated at the start meet at a certain point, travel the same distance for a while, and then break up again. In this case, the LCSS method does not consider the degree of separation between the two trajectories, but focuses only on the common trajectory, and the longer the trajectory, the more similar the trajectories are.

EDR [6, 8] is a method to calculate how much processing of the movement trajectory A should be done to match the movement trajectory B. For example, the similarity is defined as the cost of repeatedly performing insertions, deletions, or substitutions until the two match. The greater the processing cost, the lower the similarity between the two movement trajectories. Many proposals have been made regarding the definition of processing methods and costs.

4 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 so that it can evaluate the similarity of predicted and reference trajectories without wholly aligning them. The proposed GEO-BLEU should be applicable to many future studies in diverse research fields as a practical evaluation index for similarity of spatial trajectories in general.

REFERENCES

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

[2] Qiqin Cai, Lyuchao Liao, Fumin Zou, Subin Song, Jierui Liu, and Meirun Zhang. 2018. Trajectory similarity measuring with grid-based DTW. In *International Conference on Smart Vehicular Technology, Transportation, Communication and Applications*. Springer, 63–72.

[3] F.K.-P. Chan, A.W.-C. Fu, and C. Yu. 2003. Haar wavelets for efficient similarity search of time-series: with and without time warping. *IEEE Transactions on Knowledge and Data Engineering* 15, 3 (2003), 686–705. https://doi.org/10.1109/TKDE.2003.1198399

[4] Christos Faloutsos, M. Ranganathan, and Yannis Manolopoulos. 1994. Fast Subsequence Matching in Time-Series Databases. In *Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data* (Minneapolis, Minnesota, USA) *(SIGMOD '94)*. Association for Computing Machinery, New York, NY, USA, 419–429. https://doi.org/10.1145/191839.191925

[5] Toshiko Ichiye and Martin Karplus. 1991. Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Structure, Function, and Bioinformatics* 11, 3 (1991), 205–217. https://doi.org/10.1002/prot.340110305 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.340110305

[6] Jae-Gil Lee, Jiawei Han, and Kyu-Young Whang. 2007. Trajectory Clustering: A Partition-and-Group Framework. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data* (Beijing, China) *(SIGMOD '07)*. Association for Computing Machinery, New York, NY, USA, 593–604. https://doi.org/10.1145/1247480.1247546

[7] C. Myers, L. Rabiner, and A. Rosenberg. 1980. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28, 6 (1980), 623–635. https://doi.org/10.1109/TASSP.1980.1163491

[8] Costas Panagiotakis, Nikos Pelekis, and Ioannis Kopanakis. 2009. Trajectory Voting and Classification Based on Spatiotemporal Similarity in Moving Object Databases. In *Advances in Intelligent Data Analysis VIII*, Niall M. Adams, Céline Robardet, Arno Siebes, and Jean-François Boulicaut (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 131–142.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (Philadelphia, Pennsylvania) *(ACL '02)*. Association for Computational Linguistics, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[10] M. B. Priestley. 1980. STATE-DEPENDENT MODELS: A GENERAL APPROACH TO NON-LINEAR TIME SERIES ANALYSIS. *Journal of Time Series Analysis* 1, 1 (1980), 47–71. https://doi.org/10.1111/j.1467-9892.1980.tb00300.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9892.1980.tb00300.x

[11] Mark T. Robinson. 1990. The temporal development of collision cascades in the binary-collision approximation. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* 48, 1 (1990), 408–413. https://doi.org/10.1016/0168-583X(90)90150-S

[12] H. Sakoe and S. Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26, 1 (1978), 43–49. https://doi.org/10.1109/TASSP.1978.1163055

[13] Stan Salvador and Philip Chan. 2007. Toward Accurate Dynamic Time Warping in Linear Time and Space. *Intell. Data Anal.* 11, 5 (oct 2007), 561–580.

[14] Christian Schreckenberger, Simon Beckmann, and Christian Bartelt. 2018. Next Place Prediction: A Systematic Literature Review. In *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Prediction of Human Mobility* (Seattle, WA, USA) *(PredictGIS 2018)*. Association for Computing Machinery, New York, NY, USA, 37–45. https://doi.org/10.1145/3283590.3283596

[15] Pavel Senin. 2008. *Dynamic Time Warping Algorithm Review*. Technical Report. University of Hawaii at Manoa.

[16] Han Su, Shuncheng Liu, Bolong Zheng, Xiaofang Zhou, and Kai Zheng. 2020. A survey of trajectory distance measures and performance evaluation. *The VLDB Journal* 29, 1 (2020), 3–32.

[17] Floris Takens. 1980. Motion under the influence of a strong constraining force. In *Global Theory of Dynamical Systems*, Zbigniew Nitecki and Clark Robinson (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 425–445.

[18] T. K. Vintsyuk. 1968. Speech discrimination by dynamic programming. *Cybernetics* 4 (1968), 52–57. https://doi.org/10.1007/BF01074755