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

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.

• 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
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 \( i_1 \in [1:M], j_1 \in [1:N] \) 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 \( i_1 \leq i_{l+1} \) and \( j_1 \leq j_{l+1} \) 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 \( d(x_{i_l}, y_{j_l}) \):

\[
\text{cost}(P) = \sum_{l=1}^{L} d(x_{i_l}, y_{j_l})
\]

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 \( P \):

\[
\text{DTW} = \min_{P} \text{cost}(P).
\]

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:

\[
P_n = \frac{\sum_{S \in C} \sum_{n\text{-gram} \in S} \text{Count}_{\text{matched}}(n\text{-gram})}{\sum_{S \in C} \sum_{n\text{-gram} \in S} \text{Count}(n\text{-gram'})}
\]

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

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

where \( c = \text{the sum of the candidates'} \text{ lengths, and } r = \text{that of the references.} \)

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

\[
\text{BLEU} = BP \cdot \exp \left( \sum_{n=1}^{N} w_n \log p_n \right)
\]

where \( w_n \) is the positive weight summing up to 1. The original work of BLEU uses \( N = 4 \) and \( w_n = \frac{1}{4} \) 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 \((q_1, \ldots, q_p)\) where each location \( q_h \) is represented as a point in two-dimensional space. In addition, we define the similarity score \( s \) of a pair of n-grams \( g = (v_1, \ldots, v_n) \) and \( g' = (w_1, \ldots, w_n) \) on the basis of proximity as follows:

\[
s(g, g') = \prod_{k=1}^{n} \exp(-\beta d(v_k, w_k))
\]

where \( \beta \) is a parameter controlling the weight of proximity.
where \(d(\cdot, \cdot)\) is the Euclidean distance between two locations, and \(\beta\) 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 \(\text{Count}_{\text{matched}}(n\text{-gram})\) 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 \(P = \{(g_{c1}, g_{r1}), \ldots, (g_{cL}, g_{rL})\}\) where \(L\) is the shorter of the candidate’s and reference’s lengths, \(g_{c1}\) is an \(n\)-gram of the candidate sequence, and \(g_{r1}\) 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

\[
q_n = \frac{\sum_{(g_{c}, g_{r}) \in P} s (g_{c}, g_{r})}{\sum_{n\text{-gram} \in S} \text{Count}(n\text{-gram})}. 
\]  

(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

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

(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 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. Other, also 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