The Language of Place: Semantic Value from Geospatial Context

Anne Cocos and **Chris Callison-Burch**
University of Pennsylvania
acocos@seas.upenn.edu
ccb@cis.upenn.edu

Abstract

There is a relationship between what we say and where we say it. Word embeddings are usually trained assuming that semantically-similar words occur within the same *textual* contexts. We investigate the extent to which semantically-similar words occur within the same *geospatial* contexts. We enrich a corpus of geolocated Twitter posts with physical data derived from Google Places and OpenStreetMap, and train word embeddings using the resulting geospatial contexts. Intrinsic evaluation of the resulting vectors shows that geographic context alone does provide useful information about semantic relatedness.

1 Introduction

Words follow geographic patterns of use. At times the relationship is obvious; we would expect to hear conversations about actors in and around a movie theater. Other times the connection between location and topic is less clear; people are more likely to tweet about something they *love* from a bar than from home, but vice versa for something they *hate*.[1] Distributional semantics is based on the theory that semantically similar words occur within the same *textual* contexts. We question the extent to which similar words occur within the same *geospatial* contexts.

Previous work validates the relationship between the content of text and its physical origin. Geographically-grounded models of language enable toponym resolution (DeLozier et al., 2015), document origin prediction, (Wing and Baldridge, 2011; Hong et al., 2012; Han et al., 2012b; Han et al., 2013; Han et al., 2014) and tracking regional variation in word use (Eisenstein et al., 2010; Eisenstein et al., 2014; Bamman et al., 2014; Huang et al., 2016). Our work differs from earlier models; rather than modeling language with respect to an absolute, physical location (like a geographic bounding box), we model language with respect to attributes describing a type of location (like *amenity:movie_theater* or *landuse:residential*). This allows us to model the impact of geospatial context independently of language and region.

We enrich a corpus of geolocated tweets with geospatial information describing the physical environment where they were posted. We use the geospatial contexts to train *geo-word embeddings* with the *skip-gram with negative sampling* (SKIPGRAM) model (Mikolov et al., 2013) as adapted to support arbitrary contexts (Levy and Goldberg, 2014). We then demonstrate the semantic value of geospatial context in two ways. First, using intrinsic methods of evaluation, we show that the resulting geo-word embeddings themselves encode information about semantic relatedness. Second, we present initial results suggesting that because the embeddings are trained with language-agnostic features, they give a potentially useful signal about bilingual translation pairs.

2 Geo-enriching Tweets

We collected 6.2 million geolocated English tweets in 20 metro areas from Jan-Mar 2016.[2] The

[1] Under our GEO30 word embeddings, the word *love* is closer to the context *GooglePlaces:bar* than to *highway:residential*. The relationship is inverted for the word *hate*.

[2] The metro areas, chosen based on high volume of geolocated tweets collected during an initial trial period, were Atlanta, Bandung, Bogota, Buenos Aires, Chicago, Dallas, Washington DC, Houston, Istanbul, Jakarta, Los Angeles, London, Madrid, Mexico City, Miami, New York City, Philadelphia, San Francisco Bay Area, Singapore, and Toronto. We used only tweets explicitly tagged with geo-

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 99–104,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

tokens in these tweets were normalized by converting to lowercase, replacing @-mentions, numbers, and URLs with special symbols, and applying the lexical normalization dictionary of Han et al. (2012a).

To enrich our collected tweets with geospatial features, we used publicly-available geospatial data from OpenStreetMap and the Google Places API. OpenStreetMap (OSM) is a crowdsourced mapping initiative. Users provide surveyed data such as administrative boundaries, land use, and road networks in their local area. In addition to geographic coordinates, each shape in the data set includes tags describing its type and attributes, such as *shop:convenience* and *building:retail* for a convenience store. We downloaded metro extracts for our 20 cities in shapefile format. To maximize coverage, we supplemented the OSM data with Google Places data from its web API, consisting of places tagged with one or more types (i.e. *aquarium*, *ATM*, etc).

We enrich each geolocated tweet by finding the coordinates and tags for all OSM shapes and Google Places located within 50m of the tweet's coordinates. The enumerated tags become geographic contexts for training word embeddings. Figure 1 gives an example of geospatial data collected for a single tweet.

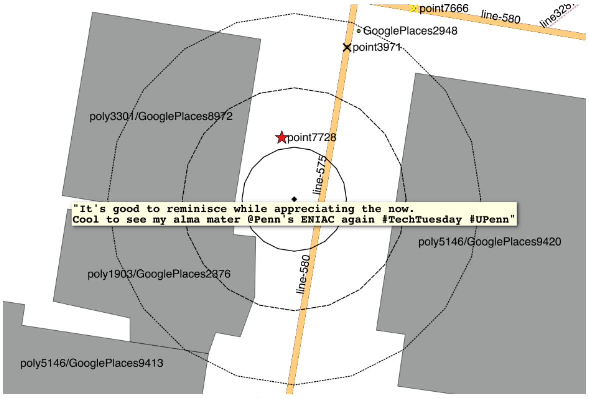
3 Geo-Word Embeddings

SKIPGRAM learns latent fixed-length vector representations v_w and v_c for each word and context in a corpus such that $v_w \cdot v_c$ is highest for frequently observed word-context pairs. Typically a word's context is modeled as a fixed-length window of words surrounding it. Levy and Goldberg (2014) generalized SKIPGRAM to accept arbitrary contexts as input. We use their software (`word2vecf`) to train word embeddings using geospatial contexts.

`word2vecf` takes a list of (word, context) pairs as input. We train 300-dimensional geo-word embeddings denoted GEO*D* – where D indicates a radius – as follows. For each length-*n* tweet, we find all shapes within D meters of its origin and enumerate the length-m list of the shapes' geographic tags. The tweet in Figure 1, for example, has $m = 10$ tags as context when training GEO30 embeddings. Under our model, each token in the tweet shares the same contexts. Thus the input to `word2vecf` for training GEO30 embeddings produced by the example tweet is an $m \times n$ list of (word, context) pairs:

```
(it’s, route:bus),
(good, route:bus),
...
(#TechTuesday, poi:marker),
(#UPenn, poi:marker)
```

The mean number of tags (m) per tweet under each threshold is 12.3 (GEO15), 21.9 (GEO30), and 38.6 (GEO50). The mean number of tokens (n) per tweet is 15.7.


point7666
line-580
line328
GooglePlaces2948
point3971
poly3301/GooglePlaces8972
point7728
line-575
"It's good to reminisce while appreciating the now.
Cool to see my alma mater @Penn's ENIAC again #TechTuesday #UPenn"
poly5146/GooglePlaces9420
poly1903/GooglePlaces2376
line-580
poly5146/GooglePlaces9413

Radius (m)	Intersecting Shapes	Geographic Tags
15	line575	route:bus
	line580	highway:tertiary
30	poly1903	building:yes, GP:university
	poly3301	building:university, GP:university
	poly5146	building:university, GP:university
	point7728	tourism:information, poi:marker
50	poly5146	building:yes, GP:university
	point3971	highway:crossing
	GooglePlaces2948	GP:bus_station

Figure 1: Geoenriching an example tweet with geographic contexts at increasing radii D (meters). For each $D \in \{15, 30, 50\}$, geographic contexts include all tags belonging to shapes within D meters of the origin. In this example there are 10 tags for the tweet at $D = 30$m. *GP* denotes tags obtained via Google Places; others are from OpenStreetMap.

4 Intrinsic Evaluation

To determine the extent to which geo-word embeddings capture useful semantic information, we first evaluate their performance on three semantic relatedness and four semantic similarity benchmarks (listed in Table 1). In each case we calcu-

graphic coordinates.

Target	Most similar (Geo30)	Most similar (Text5)
baseball	#baseball, softball, marlins, nem, dodgers	softball, lacrosse, #baseball, soccer, tourney
history	natural, dinosaurs, #naturalhistorymuseum, museum, museums	#naturalhistorymuseum, smithsonian's, #museumselfie, #dinosaur, dinosaurs
#nyc	nyc, #newyorkcity, #manhattan, #ny, 🗽	#ny, #iloveny, #nyclife, #ilovenewyork, #newyorknewyork
burger	🍔, #burger, delicious, 🍟, 👌	#burger, 🍔, fries, cheeseburger, burgers
man	have, that, years, not, don't	dude, guy, woman, hugging, he
when	like, my, but, so, it's	because, whenever, that, tfw, sometimes

Table 2: Most similar words based on cosine similarity of embeddings trained using geographic contexts within a radius of 30m (Geo30) and textual contexts with a window of 5 words (Text5).

late Spearman's rank correlation between numerical human judgements of semantic similarity or relatedness for a large set of word pairs, and the cosine similarity between the same word pairs under the geo-word embedding models.

To understand the impact of geographic contexts on the embedding model, we compare Geo15, Geo30, and Geo50 geo-word embeddings to the following baselines:

Text5: Using our corpus of geolocated tweets, we train word embeddings with `word2vecf` using traditional linear bag-of-words contexts with window width 5.

Geo30+Text5: We also evaluate the impact of combining textual and geospatial contexts. We train a model over the geolocated tweets corpus using both the geospatial contexts from Geo30 and the textual contexts from Text5.

Rand30: Because our Geo*D* models assign the same geospatial contexts to every token in a tweet, we need to rule out the possibility that Geo*D* models are simply capturing relatedness between words that frequently appear in the same tweets, like *movie* and *theater*. We implement a random baseline model that captures similarities arising from tweet co-location alone. For each tweet, we enumerate the geospatial tags (i.e. contexts) for shapes within 30m of the tweet origin. Then, before feeding the $m \times n$ list of (word, context) pairs to `word2vecf` for training, we randomly map each tag type to a different tag type within the context vocabulary. For example, `route:bus` could be mapped to `amenity:bank` for input to the model. We redo the random tag mapping for each tweet. In this way, vectors for words that always appear together within tweets are trained on the same set of associated contexts. But the randomly mapped contexts do not model the geographic distribution of words.

4.1 Intrinsic Evaluation Results

Qualitatively, we find that strongly locational words, like *#nyc*, and words frequently associated with a type of place, like *burger* and *baseball*, tend to have the most semantically and topically similar neighbors (Table 2) . Function words and others with geographically independent use (i.e. *man*) have less semantically-similar neighbors.

We can also qualitatively examine the geographic context embeddings v_c output by `word2vecf`. Recall that the SkipGram objective function pushes the vectors for frequently co-occurring v_c and v_w close to one another in a shared vector space. Thus we can find the words (Table 4) and other contexts (Table 3) most closely associated with each geographic context on the basis of cosine similarity. We find qualitatively that the word-context and context-context associations make intuitive sense.

In our intrinsic evaluation (Table 1), geo-word embeddings outperformed the random baseline in six of seven benchmarks. These results are significant ($p < .01$) based on the Minimum Required Difference for Significance test of Rastogi et al. (2015). This indicates that geospatial information *does* provide some useful semantic information. However, the Geo*D* embeddings underperformed the Text5 embeddings in all cases. And although the combined Geo30+Text5 embeddings outperformed the Text5 embeddings in 2 of 3 semantic relatedness benchmarks, the results were significant only in the case of the MEN dataset ($p < .05$). This suggests, inconclusively, that geospatial contextual information may improve the semantic relatedness content of word embeddings in some cases, but that geospatial context is no substitute for textual context in capturing semantic relationships. Nevertheless, geospatial context does provide some signal for semantic relatedness that may be useful in combination with other multimodal signals. Finally, it should be noted that the Spearman correlation achieved by all models in our tests is significantly

Data Set	Data Type	Rand30	Geo15	Geo30	Geo50	Geo30+Text5	Text5	Ref
MEN	rel	0.137[2]	0.319	0.337	0.298	**0.528**[1]	0.514	(Bruni et al., 2012)
MTURK-771	rel	0.076[2]	0.224	0.225	0.206	0.357	**0.364**	(Halawi and Dror, 2012)
WS353-R	rel	0.095[2]	0.312	0.334	0.244	**0.396**	0.382	(Agirre et al., 2009)
WS353-S	sim	0.052[2]	0.314	0.275	0.249	0.525	**0.555**	(Agirre et al., 2009)
RW	sim	0.012[2]	0.176	0.167	0.167	0.323	**0.362**[1]	(Luong et al., 2013)
SCWS	sim	0.316[2]	0.392	0.383	0.385	0.470	**0.499**[1]	(Huang et al., 2012)
SimLex	sim	0.081	0.069	0.068	0.052	0.100	**0.192**[1]	(Hill et al., 2015)

[1] Indicates a significant difference between Text5 and Geo30+Text5 results ($p < 0.05$, (Rastogi et al., 2015))
[2] Indicates Rand30 results are significantly lower than any Geo or Word embedding results ($p < 0.01$, (Rastogi et al., 2015))

Table 1: We calculate the Spearman correlation between pairwise human semantic similarity (sim) and relatedness (rel) judgements, and cosine similarity of the associated word embeddings, over 7 benchmark datasets.

Geographic context	5-most-similar contexts
GP.restaurant	GP.food, GP.point_of_interest, GP.establishment, GP.cafe, GP.bar
landuse.residential	boundary.postal_code, place.neighbourhood, landuse.commercial, landuse.retail, operator.metro
amenity.place_of_worship	religion.christian, building.church, GP.place_of_worship, GP.church, religion.muslim
GP.home_goods_store	GP.furniture_store, GP.store, GP.point_of_interest, GP.establishment, GP.electrician

Table 3: Most similar contexts, based on cosine similarity of the associated Geo30 context vectors.

Geospatial context	Most similar words (Geo30)
GP.aquarium	🐋, 🐡, 🐳, #aquarium, #jellyfish
natural.peak	#hike, overlook, #hiking, coit, mulholland
amenity.museum	history, #dinosaur, #naturalhistorymuseum, american, natural
GP.bowling_alley	🎳, saray, bowling, idarts, #bowling
religion.muslim	camii, masjid, sultan, mosque, ahmed
man_made.bridge	#bridge, #manhattanbridge, #brooklynbridge, #eastriver, 🌉

Table 4: Most similar words for target contexts, based on cosine similarity of their associated Geo30 word and context vectors.

below the current state-of-the-art; this is to be expected given the relatively small size of our training corpus (approx. 400M tokens).

5 Translation Prediction

Our intrinsic evaluation established that geospatial context provides semantic information about words, but it is weaker than information provided by textual context. So a natural question to ask is whether geospatial context can be useful in any setting. One potential strength of word embeddings trained using geospatial contexts is that the features are language-independent. Thus we infer that training geo-word embeddings jointly over two languages might yield translation pairs that are close to one another in vector space. This type of model could be applicable in a low-resource language setting where large parallel texts are unavailable but geolocated text is. To test this hypothesis, we collect an additional 236k geolocated Turkish tweets and re-train Geo30, Text5, and Geo30+Text5 vectors on the larger set.

Similar to Irvine and Callison-Burch (2013), we use a supervised method to make a binary translation prediction for Turkish-English word pairs. We build a dataset of positive Turkish-English word pairs by all Turkish words in a Turkish-English dictionary (Pavlick et al., 2014) that appear in our vector vocabulary and do not translate to the same word in English (528 words in total). We add these words and their translations to our dataset as positive examples. Then, for each Turkish word in the dataset we also select a random English word and add this pair as a negative example. Our resulting data set has 1056 word pairs, 50% of which are correct translations. We split this into 80% train and 20% test examples.

We construct a logistic regression model, where the input for each word pair is the difference between its Turkish and English word vectors, $v_f - v_e$. We evaluate the results using precision, recall, and F-score of positive translation predictions.

Table 5 gives our results, which we compare to a model that makes a random guess for each word pair. Combining geographic and textual contexts to train embeddings leads to better translation performance than using textual or geospatial contexts in isolation. In particular, with a seed dictionary of just 528 Turkish words and monolingual text of just 236k tweets, our supervised method is able to predict correct translation pairs with 67.8% precision. While the not signficant under McNemar's test (p=0.07), they are suggestive that geospatial contextual information may provide a useful signal for bilingual lexicon induction when used in combination with other methods, as in Irvine and Callison-Burch (2013).

Vector	**Precision**	**Recall**	**FScore**
`Text5`	0.600	0.574	0.587
`Geo30`	0.570	0.542	0.556
`Geo30+Text5`	**0.678**	**0.588**	**0.630**
Random	0.500	0.500	0.500

Table 5: We make a binary translation prediction for Turkish-English word pairs using their embeddings in a simple logistic regression model.

6 Conclusion

Typically word embeddings are generated using the *text* surrounding a word as context from which to derive semantic information. We explored what happens when we use the *geospatial* context – information about the physical location where text originates – instead. Intrinsic evaluation of word embeddings trained over a set of geolocated Twitter data, using geospatial information derived from OpenStreetMap and the Google Places API as context, indicated that the geospatial context does encode information about semantic relatedness.

We also suggested an extrinsic evaluation method for *geo-word embeddings*: predicting translation pairs without bilingual parallel corpora. Our experiments suggested that while geospatial context is not as semantically-rich as textual context, it does provide useful semantic relatedness information that may be complementary as part of a multimodal model. As future work, another extrinsic evaluation task that may be appropriate for *geo-word* embeddings is geolocation prediction.

Acknowledgments

We would like to thank our reviewers for their thoughtful suggestions. We are also grateful to the National Physical Science Consortium for partially funding this work.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

David Bamman, Chris Dyer, and Noah A. Smith. 2014. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834. Association for Computational Linguistics.

Elia Bruni, Gemma Boleda, Marco Baroni, and Khanh Nam Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145. Association for Computational Linguistics.

Grant DeLozier, Jason Baldridge, and Loretta London. 2015. Gazetteer-independent toponym resolution using geographic word profiles. In *AAAI*, pages 2382–2388.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PloS one*, 9(11):e113114.

Guy Halawi and Gideon Dror. 2012. The word relatedness MTURK-771 test collection, v1.0.

Bo Han, Paul Cook, and Timothy Baldwin. 2012a. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 421–432. Association for Computational Linguistics.

Bo Han, Paul Cook, and Timothy Baldwin. 2012b. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING 2012*, pages 1045–1062. The COLING 2012 Organizing Committee.

Bo Han, Paul Cook, and Timothy Baldwin. 2013. A stacking-based approach to twitter user geolocation prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 7–12. Association for Computational Linguistics.

Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. 2012. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st international conference on World Wide Web*, pages 769–778. ACM.

Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882. Association for Computational Linguistics.

Yuan Huang, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2016. Understanding us regional linguistic variation with twitter data analysis. *Computers, Environment and Urban Systems*, 59:244–255.

Ann Irvine and Chris Callison-Burch. 2013. Supervised bilingual lexicon induction with multiple monolingual signals. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 518–523. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308. Association for Computational Linguistics.

Thang Luong, Richard Socher, and Christopher Manning, 2013. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, chapter Better Word Representations with Recursive Neural Networks for Morphology, pages 104–113. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of amazon mechanical turk. *Transactions of the Association of Computational Linguistics*, 2:79–92.

Pushpendre Rastogi, Benjamin Van Durme, and Raman Arora. 2015. Multiview lsa: Representation learning via generalized cca. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 556–566. Association for Computational Linguistics.

Benjamin Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 955–964. Association for Computational Linguistics.