Towards Spatial Word Embeddings

Paul Mousset[1,2], Yoann Pitarch[1], and Lynda Tamine[1]

[1]IRIT, Université de Toulouse, CNRS, Toulouse, France
[2]Atos Intégration, Toulouse, France
{paul.mousset,yoann.pitarch,lynda.tamine}@irit.fr

Abstract. Leveraging textual and spatial data provided in spatio-textual objects (eg., tweets), has become increasingly important in real-world applications, favoured by the increasing rate of their availability these last decades (eg., through smartphones). In this paper, we propose a spatial retrofitting method of word embeddings that could reveal the localised similarity of word pairs as well as the diversity of their localised meanings. Experiments based on the semantic location prediction task show that our method achieves significant improvement over strong baselines.

Keywords: Word embeddings · Retrofitting · Spatial.

1 Introduction

The last decades have witnessed an impressive increase of geo-tagged content known as spatio-textual data or geo-texts. Spatio-textual data includes Places Of Interest (POI) with textual descriptions, geotagged posts (eg., tweets), geo-tagged photos with textual tags (eg., Instagram photos) and check-ins from location-based services (eg., Foursquare). The interplay between text and location provides relevant opportunities for a wide range of applications such as crisis management [11] and tourism assistance [5]. This prominence gives also rise to considerable research issues underlying the matching of spatio-textual objects which is the key step in diverse tasks such as querying geo-texts [24], location mention [6,9] and semantic location prediction [3,25]. Existing solutions for matching spatio-textual objects are mainly based on using a combination of textual and spatial features either for building scalable object representations [24] or for designing effective object-object matching models [3,25]. The goal of our work is to explore the idea of jointly leveraging spatial and textual knowledge to build enhanced representations of textual units (namely words) that could be used at either object representation and matching levels. The central thesis of our work is driven by two main intuitions: (1) co-occurrences of word pairs within spatio-textual objects reveal localised word similarities. For instance *dinosaur* and *museum* are semantically related near a natural history museum, but less related near an art museum; (2) As a corollary of intuition 1, distinct meanings of the same word could be conveyed using the spatial word distribution as source of evidence. For instance *dinosaur* can refer to a prehistoric animal or to a restaurant chain specifically in New York. Thus, we exploit the spatial distribution of words to jointly identify semantically related word pairs

as well as localised word meanings. To conceptualise our intuitions, we propose a retrofitting strategy [7,20] as means of refining pre-trained word embeddings using spatial knowledge. We empirically validate our research intuitions and then show the effectiveness of our proposed spatial word embeddings within semantic location prediction as the downstream task.

2 Preliminaries

2.1 Definitions and Intuitions

Definition 1. (Spatio-textual object) A spatio-textual object o is a geotagged text (eg., a POI with a descriptive text). The geotag is represented by its coordinates *(lat, lon)* referring to the geographic location l denoted $o.l$ (eg., the physical location of a POI). We adopt a word-based vectorial representation of object o including all its textual attributes (eg., POI description) $o = [w_1^{(o)}, \ldots, w_m^{(o)}]$ where each word $w_i^{(o)}$ is drawn from a vocabulary $\mathcal{W}$.

Definition 2. (Spatial distance) The spatial distance between spatio-textual objects o_i, o_j refers to the geographic distance, under a distance metric, between locations $o_i.l$ and $o_j.l$. The spatial distance between words w_i, w_j refers to an aggregated (eg., average) spatial object-object distance over the sets of spatio-textual objects O_i,O_j they respectively belong to.

Intuition 1. Words that occur in close spatio-textual objects tend to have similar meanings. Basically, the spatially closer the words are, regarding the distance between their associated objects, the closer are their meanings (eg., intuitively *cup* is semantically closer to *football* in Europe than in the USA).

Intuition 2. Let us consider a localised meaning of a word as being represented by the set of spatially similar words with respect to *intuition 1*. A word could convey different localised meanings depending on the geographical area where it is spatially dense (eg., *football* in Europe does not refer to the same sport as in the USA).

2.2 Problem definition

Based on *intuition 1*, we conjecture that spatial signals could contribute to the building of distributed representations of word vectors. As previously suggested [7,20], one relevant way is to inject external knowledge into initial learned word embeddings. However different meanings of the same word are conflated into a single embedding [10,13]. Thus, from *intuition 2*, we build for each word a set of embedding vectors based on its occurrence statistics over the associated spatio-textual objects. Formally, given a set of word vector representations $\widehat{\mathbf{W}} = \{\widehat{\mathbf{w}}_1, \ldots, \widehat{\mathbf{w}}_n\}$, where $\widehat{\mathbf{w}}_i$ is the k-dimensional embedding vector built for target word $w_i \in \mathcal{W}$, using a standard neural language model (eg., Skip-gram model [14]), the problem is how to build for each word w_i the set of associated spatial word embeddings $\widehat{\mathbf{w}}_i^s = \{\widehat{\mathbf{w}}_{i,1}^s, \ldots, \widehat{\mathbf{w}}_{i,j}^s, \ldots, \widehat{\mathbf{w}}_{i,n_i}^s\}$. Each spatial word vector

$\widehat{\mathbf{w}}^s_{i,j}$, derived from an initial embedding $\widehat{\mathbf{w}}_i$, refers to the localised distributional representation of word w_i over a dense spatial area, and n_i is the number of distinct localised meanings of word w_i derived from its spatial distribution over the spatio-textual objects O_i it belongs to.

3 Methodology

3.1 Overview

Our algorithm for building the spatial word embeddings is described in Algorithm 1. For each word w_i, we first identify the spatio-textual objects O_i it belongs to. To identify dense spatial areas of word w_i, we perform a K-Means clustering [12]. More formally, for each word w_i, we determine n_i spatial clusters represented with their respective barycenters $\mathcal{B}_i = \{\mathcal{B}_{i,1}, \cdots, \mathcal{B}_{i,n_i}\}$, where $\mathcal{B}_{i,j}$ is the j-th barycenter of word w_i and n_i the optimal number of clusters for word w_i determined using the silhouette analysis [19]. Each barycenter $\mathcal{B}_{i,j}$ can be seen as a spatial representative of the area that gives rise to a local word meanings of word w_i represented by the distributed vector $\mathbf{w}^s_{i,j}$. We detail in the following section the key step of building the spatial embedding $\mathbf{w}^s_{i,j}$ based on a retrofitting process from word embedding $\widehat{\mathbf{w}}_i$ and considering both spatially neighbour words $W^+_{i,j}$ and distant words $W^-_{i,j}$ with respect to barycenter $\mathcal{B}_{i,j}$.

3.2 Spatially Constrained Word Embedding

Our objective here is to learn the set of spatial word embeddings $\mathbf{W}^s$. We want the inferred word vector $\mathbf{w}^s_{i,j}$ (i) to be semantically close (under a distance metric) to the associated word embedding $\widehat{\mathbf{w}}_\mathbf{i}$, (ii) to be semantically close to its spatial neighbour words $W^+_{i,j}$ and (iii) to be semantically unrelated to the spatially distant words $W^-_{i,j}$. Thus, the objective function to be minimised is given by:

Algorithm 1: Algorithm for building spatial word embeddings

Input: Vocabulary $\mathcal{W}$; Set of word embeddings $\widehat{\mathbf{W}} = \{\widehat{\mathbf{w}}_1, \ldots; \widehat{\mathbf{w}}_{|\mathcal{W}|}\}$; Set of spatio-textual objects O

Output: Set of spatial word embeddings $\mathbf{W}^s = \{\mathbf{w}^s_{1,1}, \ldots, \mathbf{w}^s_{1,n_1}, \ldots, \mathbf{w}^s_{|\mathcal{W}|,n_k}\}$

for $i \in \{1, .., |\mathcal{W}|\}$ **do**

1 O_i=ExtractObjects(w_i, O)

2 SpatialClustering($O_i, \mathcal{B}_i, n_i$)

end

repeat

for $i \in \{1, .., |\mathcal{W}|\}$ **do**

for $j \in \{1, .., n_i\}$ **do**

3 $W^+_{i,j}$=Neighbours($w_i, \mathcal{B}_{i,j}$)

4 $W^-_{i,j}$=Distant($w_i, \mathcal{B}_{i,j}$)

5 $\mathbf{w}^s_{ij}$= Retrofit ($\widehat{w}_i, W^+_{i,j}, W^-_{i,j}$) (see Sect. 4.2)

end

end

until *Convergence*;

$$\Psi(\mathbf{W}^s) = \sum_{i=1}^{|\mathcal{W}|} \sum_{j=1}^{n_i} \left[\alpha \; d(\mathbf{w}^s_{i,j}, \widehat{\mathbf{w}}_i) + \beta \sum_{w_k \in W^+_{i,j}} d(\mathbf{w}^s_{i,j}, \widehat{\mathbf{w}}_k) + \gamma \sum_{w_k \in W^-_{i,j}} 1 - d(\mathbf{w}^s_{i,j}, \widehat{\mathbf{w}}_k) \right]$$

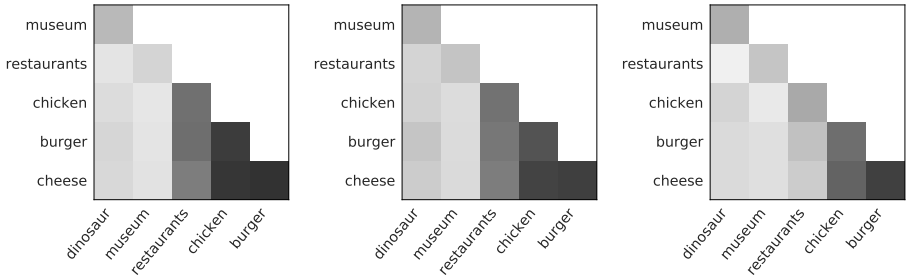
where $d(w_i, w_j) = 1 - sim(w_i, w_j)$ is a distance derived from a similarity measure (eg., cosinus), $\mathbf{W^+_{i,j}}$ (resp. $\mathbf{W^-_{i,j}}$) is the set of words spatially **close to** (resp. **distant from**) the word $w_{i,j}$, ie., words within (resp. beyond) a radius r^+ (resp. r^-) around its barycenter $\mathcal{B}_{i,j}$, and α, β, $\gamma \geq 0$ are hyperparameters that control the relative importance of each term. In our experimental setting, r^+ and r^- are set to 100 and 500 meters and $\alpha = \beta = \gamma = 1$.

4 Evaluation

4.1 Experimental Setup

Evaluation Task and Dataset. We consider the *semantic location prediction task* [3,25]. Given the tweet t, the task consists in identifying, if any, the POI p that the tweet t semantically focuses on (ie., reviews about). Formally, semantic location identifies a single POI p which is the topmost $p* \in \mathcal{P}$ of a ranked list of candidate POIs returned by a semantic matching function. We employ a *dataset* of English geotagged tweets released by Zhao et al. [25]. The dataset, consists of 74K POI-related tweets, collected from 09.2010 to 01.2015 in New York (NY) and Singapore (SG). Using the Foursquare API, we collected 800K POIs located in NY and SG cities including user-published reviews. The entire dataset consists of 238,369 distinct words, on which we applied K-Means clustering (see Sect. 3.1). As result of clustering, we found 630,732 spatial word clusters with around 2.6 local word meaning $w^s_{i,j}$ created per word w_i. We notice that 166,139 (69.7%) words have only one local meaning.

Baselines, Scenarios and Metrics. We compare our approach with a set of stat-of-the-art matching baseline models: (1) Dist [4]: the Haversine distance Tweet-POI; (2) BM25 [18]; (3) Class [25]: a POI ranking model that combines spatial distance with a text-based language model. To evaluate the effectiveness of our approach, we inject the embedding into the Class model as follows: (a) Class-Match (CM): we compute the cosine similarity of a pair (*t,p*) instead of the language model score. (b) Class-Expand (CE): we expand the tweet with the top likely similar words following the approach proposed by Zamani and Croft [23]. For the two above scenarios we consider either the traditional or the spatial word embeddings. Practically, for scenarios using spatial word embeddings, we use the closest local word $w^{(t)}_i$ (resp. $w^{(p)}_i$) by minimising the Haversine distance between tweet (resp. POI) location $t.l$ (resp. $p.l$) and word barycenters $\mathcal{B}_{i,j}$. We exploit two well-known evaluation metrics, namely *Acc@k* [17] and *Mean Reciprocal Rank* (*MRR*) [2]. Given the semantic location task description, it is worth to mention that low values of k are particularly considered.



(a) WE SIM (b) WE D-SIM (NY) (c) WE D-SIM (SG)

Fig. 1. Cosine similarities of traditional WE SIM (a), WE SIM damped by word-word barycenter distances in NY dataset (b) and in SG dataset (c)

4.2 Analysis of spatial driven word similarities

To validate the intuitions presented in Sect. 2.1, we first build as shown in Fig. 1, the heat-map of the similarity values between the embedding vectors of a sample of insightful words where the darker the cell, the more similar the pair of words. To exhibit the localised meanings of the words, we partition the dataset in two distinct subsets depending on the city the tweets were emitted from (ie., either in NY or SG). For each subset, cosine similarities are then damped by a spatial factor $f_s(w_i, w_j)$ which conveys how spatially close are the word w_i and w_j. Formally, $f_s(w_i, w_j)$ is defined as $f_s(w_i, w_j) = \exp\{-\frac{dist(\mathcal{B}_i, \mathcal{B}_j) - \mu}{\sigma}\}$ where $dist(\mathcal{B}_i, \mathcal{B}_j)$ is the Haversine distance between the barycenters of w_i and w_j and μ (resp. σ) is the average distance (resp. standard deviation) between all word pairs that describe the POIs located in the city. For simplicity purposes, we consider one barycenter per word for each subset. The heat-map of these weighted matrices are shown in Fig. 1b and Fig. 1c for NY and SG respectively. We can see for instance, that the cell (*restaurants*, *dinosaur*) is darker in Fig. 1b than in Fig. 1a while the cell is lighter in Fig. 1c than in Fig. 1a for the same word pair. Generally speaking, there is no objective obvious reason about why the words *restaurants* and *dinosaur* should be related to each other, as outlined by the similarity of their word embeddings in Fig. 1a. However, some restaurants in NY are named *Dinosaur Bar-B-Que* leading to an over-representativeness of tweets where these two terms co-occur in NY, leading to a local stronger semantic relation within this word pair in NY as revealed by Fig. 1b. This fits with our *intuition 1*. Besides, cross-looking at Fig. 1a and its spatial variants Fig. 1b and Fig. 1c provides some clues on why our *intuition 2* is well-founded. Indeed, we can see that words *dinosaur* and *museum* are similar regardless of the location. By relating this observation with the previous one, we can infer that *dinosaur* could refer to both *museum* and *restaurant* specifically in NY as revealed by the strength of its similarity with words such as *burger* and *cheese* in Fig. 1b which is clearly less pronounced in Fig. 1c.

		MRR		***Acc@1***		***Acc@5***	
		Value	R-Chg	Value	R-Imp	Value	R-Imp
Dist. based	DIST	0.514	+140.7 ∗	0.430	+19.61 ∗	0.605	+15.45 ∗
Text based	BM25	0.423	+161.3 ∗	0.307	+64.68 ∗	0.668	+4.49 ∗
Text-Dist. based	CLASS	0.507	+159.9 ∗	0.401	+25.85 ∗	0.624	+11.79 ∗
Traditional	CM-$\widehat{\mathbf{W}}$	0.521	+128.0 ∗	0.413	+24.52 ∗	0.640	+9.06 ∗
Embeddings	CE-$\widehat{\mathbf{W}}$	0.563	+119.0 ∗	0.470	+9.41 ∗	0.659	+5.94 ∗
Spatial	CM-$\mathbf{W}^s$	0.577	+128.2 ∗	0.489	+5.05 ∗	0.675	+3.36 ∗
Embeddings	CE-$\mathbf{W}^s$	0.604	–	0.515	–	0.698	–

Table 1. Effectiveness evaluation. R-Chg: CE-$\mathbf{W}^s$ relative changes. R-Imp: CE-$\mathbf{W}^s$ relative improvements. Significant Student's t-test $* : p < 0.05$.

4.3 Effectiveness

Table 1 summarises the effectiveness results obtained based on the semantic location prediction task. We compute relative changes (R-Chg) using the ratio of the geometric means of the MRR and compute the relative improvements suited for non aggregated measures for $Acc@k$. Overall, we can see that the scenarios involving matching with spatial embeddings (CM-$\mathbf{W}^s$ and CE-$\mathbf{W}^s$) significantly overpass all the compared models. For instance, CE-$\mathbf{W}^s$ displays better results in terms of MRR with relative changes ranging between 140.7% and 161.3% compared to DIST, BM25 and CLASS models. More precisely, CE-$\mathbf{W}^s$ allows a more effective mapping tweet-POI: more than 48% of the tweets are associated with the relevant POI based on the top-1 result, against 43% for DIST. In addition, we can observe that while injecting embeddings (either traditional or spatial) allows to improve the effectiveness of the CLASS model, the spatial embeddings allow the achievement of significant better performance. For instance, the MRR of the scenario CE significantly increases by 119%. Specifically looking at the two scenarios involving spatial embeddings, we can notice that CE-$\mathbf{W}^s$ improves MRR by 128.2% and $Acc@1$ by 5.05% compared to CM-$\widehat{\mathbf{W}}$. These results could be explained by the approach used to inject the embeddings. While in CE-$\mathbf{W}^s$, spatial embedding vectors are intrinsically used to expand the tweet description before the matching, they are rather used in the scenario CM-$\widehat{\mathbf{W}}$ to build tweet and POI embeddings using an IDF weighted average of embeddings which might generate biases in their representations. This observation clearly shows the positive impact of the intrinsic use of the spatial embeddings.

5 Related Work

A standard approach for improving traditional word embeddings is to inject external knowledge, mainly lexical resource constraints, using either an *online* or *offline* approach [14,16]. The online approach exploits external knowledge during the learning step [8,21,22]. For instance, Yu et al. and Xu et al. [21,22]

propose the RCM model which extends the skip-gram objective function with semantic relation between word pairs, as provided by a lexical resource, based on the assumption that related words yield similar contexts. The offline approach, also called *retrofitting*, uses external resources outside the learning step [7,15,20]. For instance, Faruqui et al. [7] propose a method for refining vector space representations by favouring related words, as provided by a lexical resource (eg., *WordNet, FramNet*), to have similar vector representations. To the best of our knowledge, our work is the first attempt for retrofitting word embeddings using spatial knowledge. To tackle the meaning conflation deficiency issue of word embeddings [1,10,13], the general approach is to jointly learn the words and their senses. For instance, Iacobacci et al. [10] first disambiguate words using the *Babelfy* resource, and then revise the continuous bag of words (CBOW) objective function to learn both word and sense embeddings.

6 Conclusion

In this paper, we introduced spatial word embeddings as a result of *retrofitting* traditional word embeddings. The retrofitting method leverages spatial knowledge toward revealing localised semantic similarities of word pairs, as well as localised meanings of words. The experimental evaluation shows that our proposed method successfully refines pre-trained word embeddings and allows achieving significant results over the semantic location prediction task. As future work, we plan to evaluate the effectiveness of our proposed spatial word embeddings within other location-sensitive tasks including spatial summarization of streaming objects such as tweets.

Acknowledgments

This research was supported by IRIT and ATOS Intégration research program under ANRT CIFRE grant agreement #2016/403.

References

1. Cheng, J., Wang, Z., Wen, J.R., Yan, J., Chen, Z.: Contextual text understanding in distributional semantic space. In: Proceedings of CIKM 2015. pp. 133–142
2. Craswell, N.: Mean reciprocal rank. Encyclopedia of Database Systems pp. 1703–1703 (2009)
3. Dalvi, N., Kumar, R., Pang, B., Tomkins, A.: A translation model for matching reviews to objects. In: Proceedings of CIKM 2009. pp. 167–176
4. De Smith, M., Goodchild, M.F.: Geospatial analysis: a comprehensive guide to principles, techniques and software tools (2007)
5. Deveaud, R., Albakour, M.D., Macdonald, C., Ounis, I.: Experiments with a venue-centric model for personalisedand time-aware venue suggestion. In: Proceedings of CIKM 2015. pp. 53–62

6. Fang, Y., Chang, M.W.: Entity linking on microblogs with spatial and temporal signals. Transactions of the Association for Computational Linguistics **2**, 259–272 (2014)
7. Faruqui, M., Dodge, J., Jauhar, S.K., Dyer, C., Hovy, E., Smith, N.A.: Retrofitting word vectors to semantic lexicons. In: Proceedings of NAACL 2015. pp. 1606–1615
8. Glavaš, G., Vulić, I.: Explicit retrofitting of distributional word vectors. In: Proceedings of ACL 2018. pp. 34–45
9. Han, J., Sun, A., Cong, G., Zhao, W.X., Ji, Z., Phan, M.C.: Linking fine-grained locations in user comments. Transactions on Knowledge and Data Engineering **30**(1), 59–72 (2018)
10. Iacobacci, I., Pilehvar, M.T., Navigli, R.: Sensembed: Learning sense embeddings for word and relational similarity. In: Proceedings of ACL and IJCNLP 2017. pp. 95–105
11. Imran, M., Castillo, C., Diaz, F., Vieweg, S.: Processing social media messages in mass emergency: A survey. ACM Computing Surveys **47**(4), 67:1–67:38 (Jun 2015)
12. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of BSMSP 1967. pp. 281–297
13. Mancini, M., Camacho-Collados, J., Iacobacci, I., Navigli, R.: Embedding words and senses together via joint knowledge-enhanced training. In: Proceedings of CoNLL 2017. pp. 100–111
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of NIPS 2013. pp. 3111–3119
15. Mrkšić, N., Séaghdha, D.O., Thomson, B., Gašić, M., Rojas-Barahona, L., Su, P.H., Vandyke, D., Wen, T.H., Young, S.: Counter-fitting word vectors to linguistic constraints. arXiv preprint (2016)
16. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of EMNLP 2014. pp. 1532–1543
17. Powers, D.M.: Evaluation: from precision, recall and f-measure to roc, informedness, markedness & correlation. Journal of Machine Learning Technologies pp. 37–63 (2011)
18. Robertson, S.E., Jones, K.S.: Relevance weighting of search terms. Journal of the American Society for Information science **27**(3), 129–146 (1976)
19. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
20. Vulić, I., Mrkšić, N.: Specialising word vectors for lexical entailment. In: Proceedings of NAACL-HLT 2018. pp. 1134–1145
21. Xu, C., Bai, Y., Bian, J., Gao, B., Wang, G., Liu, X., Liu, T.Y.: Rc-net: A general framework for incorporating knowledge into word representations. In: Proceedings of CIKM 2014. pp. 1219–1228
22. Yu, M., Dredze, M.: Improving lexical embeddings with semantic knowledge. In: Proceedings of ACL 2014. pp. 545–550
23. Zamani, H., Croft, W.B.: Estimating embedding vectors for queries. In: Proceedings of ICTIR 2016. pp. 123–132
24. Zhang, D., Chan, C.Y., Tan, K.L.: Processing spatial keyword query as a top-k aggregation query. In: Proceedings of SIGIR 2014. pp. 355–364
25. Zhao, K., Cong, G., Sun, A.: Annotating points of interest with geo-tagged tweets. In: Proceedings of CIKM 2016. pp. 417–426