Aligning Word Embeddings of Different Languages within a Shared Embedding Space

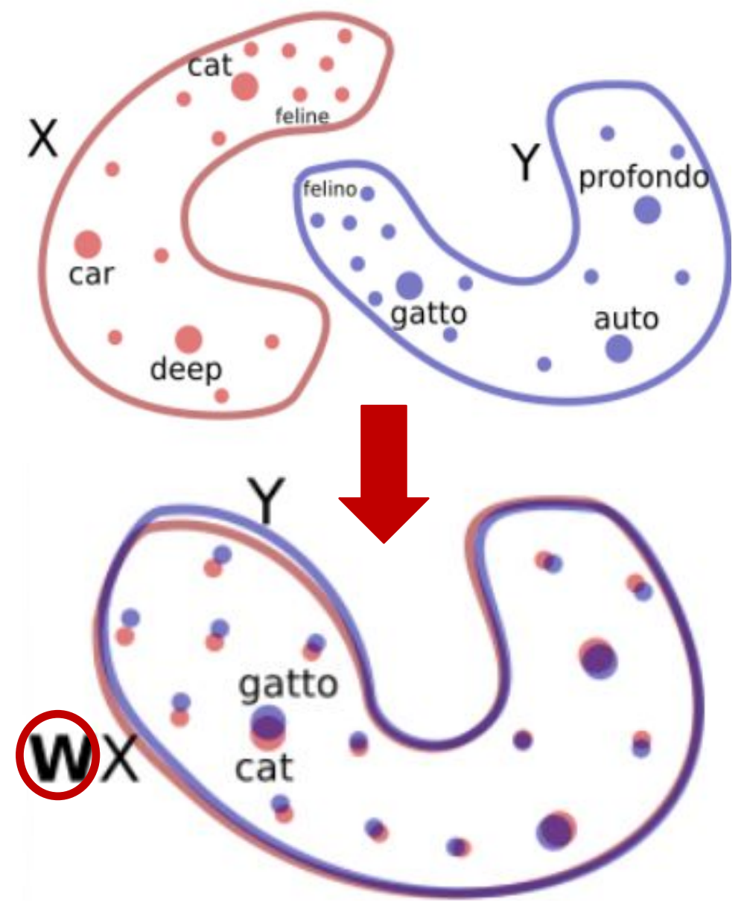
Barun Patra, Maximilian Sieb, Hima Tammineedi

Carnegie Mellon University, School of Computer Science

Introduction

Can we formulate a **convex optimization problem** for learning mappings between **word embedding spaces** of different languages?

How can we leverage knowledge of multiple dictionaries by mapping into a **shared embedding space**?



Background

Geometry aware Multilingual Mapping (GeoMM)

We follow the work of Jawanpuria et al. (2018). Given word embedding spaces $\mathcal{X}_s, \mathcal{X}_t$ we construct a mapping matrix $Y_{st}, Y_{st} \in \mathbb{R}^{|X_s| \times |X_t|}$, such that $y_{i,j} \in Y = 1$ if $x_i \in \mathcal{X}_s$ is aligned with $x_j \in \mathcal{X}_t$, else 0. We then optimize the following loss function:

$$\max_{U_1, U_2 \cdots U_n \in \mathbb{O}^d, B \succ 0} \lambda ||B||_F^2 + \sum_{l_i, l_j \in \mathcal{L}} ||X_i U_i B U_j^T X_j^T - Y_{ij}||_F^2$$

Where $\mathbb{O}^d$ denotes the family of orthogonal matrices, and $B \succ 0$ is a learned distance metric, taking into account feature correlations.

The multilingual extension for the same is pretty straightforward then. Given $\mathcal{L} = \{l_1 \cdots l_n\}$, a set of languages, we then learn a collection of U_i matrices, one for each language, to transform them into a common embedding space, and then learn a $B \succ 0$ distance matrix, to learn the distance in the common embedding space. Specifically, the following objective function is optimized:

$$\max_{U_1, U_2 \cdots U_n \in \mathbb{O}^d, B \succ 0} \lambda ||B||_F^2 + \sum_{l_i, l_j \in \mathcal{L}} ||X_i U_i B U_j^T X_j^T - Y_{ij}||_F^2$$

Relaxed Cross-domain Similarity Local Scaling (RCSLS)

We follow the work of Joulin et al. (2018). Their work proposes a new relaxed version of the cross-domain similarity local scaling (CSLS) loss originally proposed by Lample et al. (2018) for learning a mapping between two (normalized) embedding spaces X and Y:

$$CSLS(x, y) = -2\cos(x, y) + \frac{1}{k} \sum_{y' \in \mathcal{N}_Y(x)} \cos(x, y') + \frac{1}{k} \sum_{x' \in \mathcal{N}_X(y)} \cos(x', y)$$

$N_X(y)$ indicates the nearest neighbors of y taken from the space Wx (where W is a mapping matrix), and cos is the cosine similarity metric.
This loss intuitively minimizes the cosine distance between aligned words, while penalizing the distance between a word and its nearest neighbors.

Joulin et al. (2018) propose the following Relaxed version of the CSLS loss, dubbed RCSLS, by following previous work and constraining W to be orthogonal and assuming that the word vectors are ℓ_2-normalized. Note that the loss here is formulated in terms of all word pairs:

$$RCSLS(\mathbf{x}, \mathbf{y}) = \min_{W \in \mathcal{O}^d} \frac{1}{n} \sum_{i=1}^{n} -2x_i^T W^T y_i + \frac{1}{k} \sum_{y_j \in \mathcal{N}_Y(Wx_i)} x_i^T W^T y_j + \frac{1}{k} \sum_{Wx_j \in \mathcal{N}_X(y_i)} x_j^T W^T y_i$$

We can now relax the orthogonality constraint on W by using the convex hull $\mathcal{C}^d$ (the unit ball of the spectral norm).
The loss can now be rewritten as

$$RCSLS(\mathbf{x}, \mathbf{y}) = \min_{W \in \mathcal{C}^d} \frac{1}{n} \sum_{i=1}^{n} -2x_i^T W^T y_i + \frac{1}{k} \max_{S \in S_k(n)} \sum_{j \in S} x_i^T W^T y_j + \frac{1}{k} \max_{S \in S_k(n)} \sum_{j \in S} x_j^T W^T y_i$$

where $S_k(n)$ denotes the set of all subsets of $\{1, \ldots, n\}$ of size k.

Methods

Combining GeoMM and RCSLS

RCSLS has the benefits that it optimizes a loss function actually used during test time for retrieval, and this loss is convex. However, this method only works for pair-wise mappings.
GeoMM has the benefit that it maps languages into a shared embedding space. The drawbacks are that it uses different loss functions during training and test, and the loss function is not convex.

Thus, we propose to combine these two methods in order to produce a method that is convex, uses the same optimization function at both train and test times, and maps languages into a shared embedding space. We use the RCSLS loss formulation and optimize over the set of orthogonal matri U_s, U_t and the set of positive-definite matrices B:

$$\min_{U_s \in \mathcal{O}^d, U_t \in \mathcal{O}^d, B \succ 0} \frac{1}{n} \sum_{i=1}^{n} -2x_i^T U_s^T B U_t y_i + \frac{1}{k} \sum_{y_j \in \mathcal{N}_Y(x_i^T U_s^T B U_t)} x_i^T U_s^T B U_t y_j + \frac{1}{k} \sum_{x \in \mathcal{N}_X(U_s^T B U_t)} x_j^T U_s^T B U_t y_i$$

Problem Reparameterization

Consider the fact that the distance matrix $B \succ 0$ learned is symmetric. Consequently, it can be factorized as $B = Q \Lambda Q^T$, where $Q \in \mathbb{O}^d$, and Λ is a diagonal matrix with positive values. Thus, we can rewrite the objective as

$$\min_{\tilde{U}_s \in \mathcal{O}^d, \tilde{U}_t \in \mathcal{O}^d, \Lambda \succ 0} \frac{1}{n} \sum_{i=1}^{n} -2x_i^T \tilde{U}_s^T \Lambda \tilde{U}_t y_i + \frac{1}{k} \sum_{y_j \in \mathcal{N}_Y(x_i^T \tilde{U}_s^T \Lambda \tilde{U}_t)} x_i^T \tilde{U}_s^T \Lambda \tilde{U}_t y_j + \frac{1}{k} \sum_{x \in \mathcal{N}_X(\tilde{U}_s^T \Lambda \tilde{U}_t)} x_j^T \tilde{U}_s^T \Lambda \tilde{U}_t y_i$$

Estimating $\tilde{U}_s, \tilde{U}_t, \Lambda$ requires estimating $2d^2 + d$ parameters over the $3d^2$ parameters required to be estimated for the original formulation containing B. Consequently, this should make the estimation problem easier, particularly when there is less data.

Antonym Penalty

To improve the mapping between language word embeddings, we hypothesize that incorporating antonym pairs to enforce larger distances between words with opposing meanings should help.

Thus, we used a WordNet antonyms list and incorporated this into the the RCSLS loss.

$$\lambda \frac{1}{n_A} \sum_{i=1}^{n_A} 2z_i^T W^T W \tilde{z}_i$$

Here, $\{z_i\}_{i=1,\ldots,n_A}$ is the subset of all words $\{x_i\}_{i=1,\ldots,n}$ for which we have corresponding antonyms $\tilde{z}_i$ in the training set. This term can be seen as an additional regularizer to enforce a certain distance in the target domain between antonyms.

Experimental Results

GeoMM	en-es	en-fr	en-ru	en-zh
GeoMM	82.6	82.8	51.3	49.1
RCSLS	84.5	83.1	57.5	45.1
Combined (ours)	83.1	81.7	54.0	49.5

We see that our combined method generally matches or improves upon the GeoMM formulation.

The performance is worse compared to the RCSLS baseline. However, we do not expect our combined method to improve over RCSLS as in both the GeoMM and our method, we first map to a common embedding space, as opposed to the direct mapping between embedding spaces that RCSLS performs. In that sense, RCSLS directly optimizes for bilingual mappings whereas our method and GeoMM optimizes for multilingual settings.

The RCSLS task is thus generally easier, and so it should have higher performance. Our combined method therefore loses some of the accuracy benefits in exchange for theoretically better performance on multiple languages at a time.

Discussion & Future Directions

The individual baseline results of both RCSLS and GeoMM are comparable to each other. RCSLS optimizes for a direct mapping between two languages, while GeoMM maps via a shared embedding space. Therefore, the problem solved by GeoMM is harder.

We reparametrized the problem of GeoMM to decrease the number of estimated parameters and projected into the space of diagonal PSD matrices. However, the optimized parameters yield worse performance than using the original formulation.

The combined approach leads to improvements for some language pairs, but not others, compared to the individual benchmarks of RCSLS and GeoMM.

Our approach of using antonyms to instill a prior into the optimization that words with opposing meanings should be farther apart also did not yield improved results over the baselines. We suspect that the antonym regularization did not lead to improved performance because this regularization is still not enough to enforce structure in the embedding space. This regularization was meant to enforce some structure, but it was perhaps not enough.

One observation that was made about the original RCSLS paper was that the optimization performed there is unconstrained, and the authors merely stop the optimization process after some number of iterations. We believe that there is opportunity for a more principled approach here, finding explicit constraints for the optimization procedure.

References

Jawanpuria, P.; Balgovind, A.; Kunchukuttan, A.; Mishra, B. arXiv preprint arXiv:1808.08773 2018
Lample, G.; Conneau, A.; Ranzato, M.; Denoyer, L.; Jégou, H. Word translation without parallel data. International Conference on Learning Representations. 2018.
Joulin, A.; Bojanowski, P.; Mikolov, T.; Grave, E.CoRR2018, Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion