**Summary**

- A direct optimization approach to cross-lingual word embedding alignment
- The Gromov-Wasserstein distance is well-suited for this task because it:
  - Relies on relational rather than positional similarities across spaces
  - Applies to embeddings of different algorithms and dimensionality too
- Unsupervised objective strongly predictive of final accuracy

**Motivation**

- Many tasks in NLP rely on learning cross-domain correspondences
- Parallel data not always available \(\implies\) unsupervised methods
- Word-2-vec translation (bilingual lexical induction): a simple, but important litmus test
- Recent fully unsupervised methods perform on par with supervised counterparts [1, 2]
- ... but adversarial training is slow and often unstable

**Background**

**Discrete Optimal Transport**

\[
\mu = \sum_{i=1}^{m} \pi \delta_{x(i)} \quad \nu = \sum_{j=1}^{n} q \delta_{y(j)}
\]

\[C_{ij} = C(x(i), y(j))\]

- Discrete distributions: \(\pi = \sum_{i} \pi \delta_{x(i)}\), \(\nu = \sum_{j} q \delta_{y(j)}\)
- Pairwise costs: \(C_{ij} = C(x(i), y(j))\)
- Feasible couplings, \(\Gamma \in \mathbb{R}^{m,n}\) in:
  \[\Gamma \in \mathbb{R}^{m,n} \quad \Gamma \geq 0, \quad \sum_{i} \Gamma_{ij} = \pi_i, \quad \sum_{j} \Gamma_{ij} = \nu_j\]

- The problem: \(\min_{\Gamma \in \mathbb{R}^{m,n}} \sum_{i,j} C_{ij} \Gamma_{ij}\)

**Optimal Transport between Word Embeddings**

- Previous applications:
  - Word Move’s Distance [Kusner et al., 2015]: sentence similarity
  - In Word Embedding Alignment [3]
- Treat embeddings as support points of discrete distribution
  \(C_{ij} = c(x^{(i)}, y^{(j)}) = d(x^{(i)}, y^{(j)})\)
  - But this assumes the two spaces are registered
- Not true in general for word embeddings!

**The Gromov-Wasserstein Distance**

- Generalizes OT to the non-registered case
- Main idea: compare distances instead of absolute positions

\[
\begin{equation}
\begin{aligned}
C(x^{(i)}, y^{(j)}) = & \quad L(d(x^{(i)}, x^{(j)})) \\
& \quad L(d(y^{(i)}, y^{(j)}))
\end{aligned}
\end{equation}
\]

- The objective:

\[GW(C, C', p, q) = \min_{\Gamma \in \mathbb{R}^{m,n}} \sum_{i,j,k,l} L(C_{ik} C_{lj} \Gamma_{ij} \Gamma_{kl})\]

**Aligning Embedding Spaces with GW**

**Experiments**

**Training Dynamics**

- Objective closely follows the metric of interest (accuracy, not available during training)

**Translation Accuracy Results**

- TLDR: Comparable with SOTA
- Significantly (order of magnitude) faster than adversarial approaches

**The GW Linguistic Distance**

- Recall: GW problem induces a (true) metric
- Notion of semantic-syntactic ling. distance

**Discussion + Future Work**

- Speed-ups using GPU + stochastic opt
- Experiments on different embedding algorithms and dimensionality
- Extension to sentence level translation

---

**Key References**